arXiv:2604.14360v1 [cs.CR] 15 Apr 2026

Digital Guardians: The Past and The Future of Cyber-Physical Resilience

SAURABH BAGCHI*, Purdue University, United States of America
HYUNSEUNG KIM*, Purdue University, United States of America
TAREK ABDELZAHER, University of Illinois Urbana-Champaign, United States of America
HOMA ALEMZADEH, The University of Virginia, United States of America
SOMALI CHATERJI, Purdue University, United States of America
GLEN CHOU, Georgia Institute of Technology, United States of America
YUYING DUAN, University of Notre Dame, United States of America
FANXIN KONG, University of Notre Dame, United States of America
MICHAEL LEMMON, University of Notre Dame, United States of America
YIN LI, University of Wisconsin–Madison, United States of America
MENGYU LIU, University of Notre Dame, United States of America
WENHAO LUO, University of Illinois Chicago, United States of America
MEIYI MA, Vanderbilt University, United States of America
SIBIN MOHAN, George Washington University, United States of America
AYAN MUKHOPADHYAY, William & Mary, United States of America
MELKIOR ORNIK, University of Illinois Urbana-Champaign, United States of America
DIMITRA PANAGOU, University of Michigan, United States of America
KRISTIN YVONNE ROZIER, Iowa State University, United States of America
IVAN RUCHKIN, University of Florida, United States of America
HUAJIE SHAO, William & Mary, United States of America
SZE ZHENG YONG, Northeastern University, United States of America
MAJID ZAMANI, University of Colorado Boulder, United States of America
XUGUI ZHOU, Louisiana State University, United States of America

*Other than the first two authors, all other authors are listed alphabetically. Corresponding author: Saurabh Bagchi.
Authors' addresses: Saurabh Bagchi*, Purdue University, West Lafayette, IN, United States of America, sbagchi@purdue.edu; Hyunseung Kim*, Purdue University, West Lafayette, IN, United States of America, kim4061@purdue.edu; Tarek Abdelzaher, University of Illinois Urbana-Champaign, Champaign, IL, United States of America, zaher@illinois.edu; Homa Alemzadeh, The University of Virginia, Charlottesville, VA, United States of America, alemzadeh@virginia.edu; Somali Chaterji, Purdue University, West Lafayette, IN, United States of America, schaterj@purdue.edu; Glen Chou, Georgia Institute of Technology, Atlanta, GA, United States of America, chou@gatech.edu; Yuying Duan, University of Notre Dame, Notre Dame, IN, United States of America, yduan2@nd.edu; Fanxin Kong, University of Notre Dame, Notre Dame, IN, United States of America, fkong@nd.edu; Michael Lemmon, University of Notre Dame, Notre Dame, IN, United States of America, lemmon@nd.edu; Yin Li, University of Wisconsin–Madison, Madison, WI, United States of America, yin.li@wisc.edu; Mengyu Liu, University of Notre Dame, Notre Dame, IN, United States of America, mliu9@nd.edu; Wenhao Luo, University of Illinois Chicago, Chicago, IL, United States of America, wenhao@uic.edu; Meiyi Ma, Vanderbilt University, Nashville, TN, United States of America, meiyi.ma@vanderbilt.edu; Sibin Mohan, George Washington University, Washington, DC, United States of America, sibin.mohan@gwu.edu; Ayan Mukhopadhyay, William & Mary, Williamsburg, VA, United States of America, amukhopadhyay@wm.edu; Melkior Ornik, University of Illinois Urbana-Champaign, Champaign, IL, United States of America, mornik@illinois.edu; Dimitra Panagou, University of Michigan, Ann Arbor, MI, United States of America, dpanagou@umich.edu; Kristin Yvonne Rozier, Iowa State University, Ames, IA, United States of America, kyrozier@iastate.edu; Ivan Ruchkin, University of Florida, Gainesville, FL, United States of America, iruchkin@ece.ufl.edu; Huajie Shao, William & Mary, Williamsburg, VA, United States of America, hshao@wm.edu; Sze Zheng Yong, Northeastern University, Boston, MA, United States of America, s.yong@northeastern.edu; Majid Zamani, University of Colorado Boulder, Boulder, CO, United States of America, majid.zamani@colorado.edu; Xugui Zhou, Louisiana State University, Baton Rouge, LA, United States of America, xuguizhou@lsu.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Resilience in cyber-physical systems (CPS) is the fundamental ability to maintain safety and critical functionality despite adverse "perturbations," which includes security attacks, environmental disruptions, and hardware or software failures. This survey provides a comprehensive review of CPS resilience, framing the field through five interconnected themes that are required in an integrated whole to achieve real-world resilience.

The article first posits that resilience is a system-wide property emerging from interactions between hardware, software, and human users. Second, it addresses the challenges of learning-enabled CPS, which often operate in data-scarce environments characterized by imbalanced or noisy data, requiring innovative solutions like synthetic data generation and foundation model adaptation. Third, the survey examines proactive measures for resilience, which include distinctive aspects of verification, testing, and redundancy. Fourth, it explores recovery mechanisms, moving beyond traditional fault models to design "just good enough" recovery strategies that prioritize safety-critical functions during perturbations. Finally, it highlights the central role of the human, focusing on the different levels of human intervention, the necessity of trust calibration, and the requirement for explainable AI to support human-CPS teaming.

These themes are illustrated through representative application domains, primarily Connected and Autonomous Transportation Systems (CATS) and Medical CPS (MCPS). By integrating the five interconnected themes, this survey provides a systematic roadmap for achieving the resilient CPS in increasingly complex and adversarial environments.

CCS Concepts: • **Computer systems organization → Cyber-physical systems**; • **Security and privacy → Systems security**; • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: CPS reliability and security, System-wide resilience, Learning-enabled CPS, Proactive and reactive mechanisms, Human-machine teaming

ACM Reference Format:
Saurabh Bagchi*, Hyunseung Kim*, Tarek Abdelzaher, Homa Alemzadeh, Somali Chaterji, Glen Chou, Yuying Duan, Fanxin Kong, Michael Lemmon, Yin Li, Mengyu Liu, Wenhao Luo, Meiyi Ma, Sibin Mohan, Ayan Mukhopadhyay, Melkior Ornik, Dimitra Panagou, Kristin Yvonne Rozier, Ivan Ruchkin, Huajie Shao, Sze Zheng Yong, Majid Zamani, and Xugui Zhou. 2025. Digital Guardians: The Past and The Future of Cyber-Physical Resilience. *ACM Comput. Surv.* 37, 4, Article 111 (August 2025), 42 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Resilience in cyber-physical systems (CPS) is the ability to maintain critical functions and safety in the face of adverse events like security attacks, environmental disruptions, unintended human-system interactions, and natural failures in the hardware, the software, or their interfaces. Resilience goes beyond the traditional notion of detection and recovery; it focuses not just on preventing failures but also on ensuring the system can withstand, recover from, and adapt to adverse events. We use the term *"perturbations"* to include all the types of adverse events mentioned above.

In this article, we shed new light on the five key themes required to achieve CPS resilience, under practical real-world constraints. These constraints come from the fact that these systems often include many legacy elements, which cannot be "upgraded" for reliability or security. Further, CPS will have interactions with human users at different cognitive and application-specific skill levels; as such, it is impossible to instantiate practical systems for all types of interactions. Finally, there might exist policy and administrative restrictions on what is allowable for making CPS resilient. For example, different stakeholders may own different parts of a CPS — think of any of a variety of large-scale CPS that surround us today, like energy distribution, transportation, and industrial supply chain. The different stakeholders may have only partially aligned incentives, and there are regulatory and competitive factors that prevent cooperation.

© 2025 Association for Computing Machinery.
Manuscript submitted to ACM CSUR

Under this set of real-world constraints, we posit that there are **five key themes** for CPS resilience. Individually, elements within each theme have been discussed in prior literature; what this article brings is a systematic way of tying each theme to the system goal of CPS resilience and the interplay among these aspects.

(1) *Resilience is a system-wide property.* This implies that resilience is *not* an individual-component property; rather, resilience needs to be ensured through interactions among hardware-software components, as well as between such components and the human users of the system. On the flip side, a reduction in system resilience can happen due to vulnerable interactions. This shines the spotlight on the robustness of decision and component boundaries for ensuring system resilience.
(2) *Data for learning in CPS is a challenge.* Data in any learning-enabled system is a key ingredient for its resilience. Distinctively, data in CPS is expensive to collect and expensive to label. Thus, for the foreseeable future, we expect to be in a regime of low availability of data: it will be hugely disbalanced, limited for failures, and barely available if at all for catastrophic failures. This clues us into solution directions, e.g., creating synthetic data and learning from near-failures. Also, data will have various modalities and will need to be integrated for integrated situational awareness, leading to resilience.
(3) *Verification, testing, and redundancy for CPS resilience.* This theme is well understood in the context of the resilience of any system. For CPS, this theme brings out some distinctive aspects. For one, timeliness must be part of the equation, as functionality must be provided within the required time bounds. Achieving high coverage in testing CPS is a challenge when there is close interaction between the cyber and the physical elements. A simple but useful imperative is to achieve high coverage for the combinations of the two kinds of elements.
(4) *Role of recovery.* This is a complement to the above theme and focuses on mitigation and recovery. Resilience incorporates the ability to handle perturbations beyond the pre-specified, and therefore designed for, fault models. The key question here is how do we recover from such situations, rarely to a fully functional state. Rather, we need to understand and design for "just good enough" recovery, one that keeps the security- or safety-critical functionality of the system available.
(5) *Role of the human.* Frequently mentioned in design documents and papers on CPS resilience, a gap often exists between how the human role is considered in the design and implemented in the deployment. A deep consideration of this factor opens up areas of inquiry. For example, the fact that human users will behave only in a partially rational manner must factor into the fault model. As another example, explainability and trustworthiness of the algorithms become crucial, with explainability being a mix of ante-hoc and post-hoc reasoning.

A high-level overview of how the different themes interact among themselves is shown in Figure 1. This shows that the proactive, learning-enabled CPS, and mitigation and recovery themes work in an integrated manner to lead to CPS resilience being a system-wide property, rather than a property for individual or sets of components. Humans interface with the overall system at different spatial and temporal granularities, through design, deployment, operation, and getting the results and providing feedback.

Exemplars of CPS Resilience in the World.

The positive examples of resilience in CPS often show up through the Sherlockian syndrome of "the dog that did not bark." For example, the internet infrastructure in Japan proved to be largely resilient through the Great East Japan Earthquake of 2011. This was attributed to proactive provisioning of redundancy, such as through backup network lines, as well as to the rapid restoration effort by Japanese telecom companies like NTT. For a more recent case, and one with mixed examples, consider Ukraine's power grid in the ongoing Russo-Ukrainian war. Initially, the power grid withstood

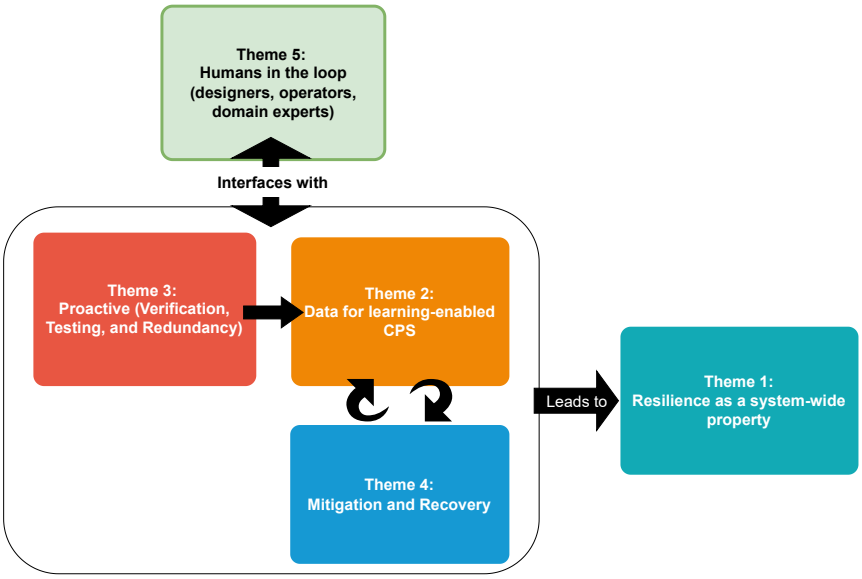

Fig. 1: High-level view of the interplay between the various themes developed in this article leading to resilient CPS

the destruction surprisingly well, preventing large-scale blackouts. The nation had made dramatic improvements to the resilience of its power grid after a malware attack had severely disrupted the power grid in 2015. However, under more intense attacks in 2024 and 2025, which have been informed by previous defenses, the Ukrainian power grid has now been brought to its knees, operating at only about a third of its pre-invasion generation capacity. This is caused by coordinated attacks against multiple parts of its infrastructure — physical (natural gas, hydroelectric, and thermal power) and cyber (large-scale cyberattacks, coordinated with the physical attacks).

Sometimes, a successful case of resilience is harder to spot due to the far more newsworthy headlines of the devastating impact of failures. A case in point is the ALERTCalifornia wildfire detection system, a sophisticated network of over 1,100 high-definition, pan-tilt-zoom cameras equipped with AI software. In the first two months of its deployment in 2023, the system had correctly identified 77 fires before any 911 calls came in. A notable success was the River Fire in San Diego County in August 2024, where the AI software classified the smoke plume as dangerous and alerted authorities nine minutes before the first public report. This early warning allowed for a rapid aerial and ground response that kept the fire to just 54 acres. This is a textbook example of the three constituent elements coming together to achieve CPS resilience — cyber (the software, the networked aspect), physical (the cameras mounted on high places), and human (the personnel actually performing the response). However, the LA wildfire in January 2025 caused devastating loss of property ($250B) and lives (31 confirmed deaths), and such failures of CPS resilience and infrastructure gathered far more public attention.

2 Representative Application Domains

Cyber-physical systems (CPS) are increasingly being used for "smart" infrastructure systems providing critical services to human society. Examples are readily found in smart sewer systems [146], medical CPS (MCPS) [110], smart power grids [252] and smart highway systems [152]. For infrastructure CPS, resilience moves beyond concerns over component

reliability and security to concerns over the system's capacity to serve the target parts of the society, which is almost always at a large scale. Next, we are going to discuss the CPS resilience challenges in two representative application domains — Connected and Autonomous Transportation System (CATS) and Medical CPS. In the end, after the description of the five themes, we will bring back how these themes collectively largely address these challenges.

2.1 Connected and Autonomous Transportation Systems (CATS)

The safety and reliability of autonomous vehicles can be significantly enhanced with *connected autonomy*, in which driving decisions are made with the assistance of, or in cooperation with, other vehicles or infrastructure elements (road-side sensors, edge computing servers). This can reduce cost and improve safety by overcoming limitations of in-vehicle autonomy [189]. An autonomous vehicle has three components: Perception, Planning, and Control. *Perception* processes the output of depth perception sensors like LiDAR or stereo cameras to build a time-varying model of the physical environment around the vehicle, updated in real-time. *Planning* uses this model to devise a strategy for reaching the destination at three different timescales — duration of the trip, seconds for nearby environment, and sub-second for immediate environment. *Control* takes this planned trajectory and translates it into actions for the vehicle's actuators, such as steering, acceleration, and braking (at timescales of a few milliseconds).

As of today, ***all three components (perception, planning, and control) operate onboard the autonomous vehicle itself***. A connected autonomous transportation system (CATS) is a class of CPS in which some of these components reside in other vehicles, or in the infrastructure. This system can enhance the functionality and safety of autonomous vehicles in several ways. For example, in ***V2V Cooperative Perception*** each vehicle shares the processed data from its own sensors with other vehicles [184, 263], increasing safety by overcoming sensor line-of-sight limitations. In ***Infrastructure-aided Perception***, an array of roadside sensors [76, 204] can track objects, use an edge computing platform to compute a time-varying model of the environment, and deliver it wirelessly to each vehicle. Finally, in ***Infrastructure-aided Planning***, the edge computing platform can plan paths and trajectories on behalf of the vehicle, and deliver these wirelessly to each vehicle, where the on-board control component implements them. Even in a highly connected CATS setting, each vehicle still requires a dependable local perception stack because wireless connectivity can be intermittent, delayed, or unavailable. Recent adaptive onboard perception systems, including 2D approaches such as LiteReconfig [239] and 3D approaches such as Agile3D [240], illustrate this need by dynamically reconfiguring perception pipelines under changing scene complexity, latency constraints, and resource contention. Thus, adaptive local perception complements V2V and infrastructure-assisted autonomy by providing robust fallback and graceful degradation when external support is unreliable. However, the Control component *cannot be offloaded to the infrastructure*, since control tasks, *e.g.*, steering, acceleration, and braking, operate require millisecond-level response times that can only be achieved with an onboard control system.

This vision of CATS raises several challenges for reliability and security of CPS, some quite distinctive to this application domain. This vision relies on wireless networks, short range or long range, to be provisioned and operated dependably. It relies on trust at multiple end points, both static endpoints (such as roadside units or RSUs) and mobile endpoints (obviously, other vehicles). With respect to any vehicle, the connectivity graph is also dynamic highlighting another challenge for reliability and security.

2.2 Medical Cyber-Physical Systems (MCPS)

Advances in sensing, computing, networking, and machine learning technologies have resulted in the ubiquitous deployment of medical cyber-physical systems in various clinical and personalized settings, ranging from implantable

and wearable devices such as pacemakers, physiological monitors, insulin pumps, and artificial pancreas systems (APS), to smart hospital devices such as infusion pumps, intensive care units (ICUs) monitors, and surgical robots, to image-guided surgery and radiation therapy systems.

Within an emergency room or ICU, MCPS are essentially networked embedded systems that must reliably monitor patient vitals in a highly dynamic, uncertain, and unsecured environment. However, with the aging US population, MCPS are increasingly being used to connect an outpatient's embedded medical devices to a centralized medical server. The sum of all these interconnected outpatient devices and servers forms an enterprise-scale MCPS whose main function is to improve the overall health of the served populations. Although component reliability and security are important issues, the overall effectiveness of the system now depends on the extent to which the MCPS improves general public health. This shifts the mission from ensuring favorable individual health outcomes to ensuring the average outcome for all individuals in the community. The success of that system-wide MCPS mission depends on the quality of the data used to train diagnostic models running on these embedded devices; data that are subject to abrupt distribution shifts triggered by internal and exogenous forces. These distribution shifts in MCPS training data present enormous challenges to enterprise-scale MCPS resilience that can only be met if the system can detect and adaptively react to such shifts.

Medical Cyber-Physical Systems (MCPS) [111] used in remote patient monitoring (RPM) provide a compelling example of CPS's requiring system-wide resilience. RPM or Telemonitoring [167] uses wearable digital devices to record a patient's physiological function outside of a traditional hospital setting. The data generated by these devices is transmitted to hospital servers. Over the past decade [219], this data has been used by data analytic models to notify the user and care provider regarding the need for follow-up care or critical medical interventions. RPM is used for diabetes management [196], post-operative cardiac care [16] , and end-of-life palliative care [163]. In these settings, MCPS function is evaluated through metrics measuring its impact on overall health outcomes and overall healthcare costs. These metrics are not a sole function of device reliability and accuracy, they also hinge on the community-wide impact such systems have on healthcare outcomes and costs. In viewing MCPS as providing a community-wide service, the resilience of such systems must embrace a range of issues including device reliability, data integrity and security, and the robustness of data analytic models.

The growing use of personal devices (smartwatches, smart rings) for health monitoring opens a new aspect of MCPS resilience. With the data collected from these personal devices, MCPS take on the additional role of monitoring and maintaining community public health in a way that is resilient to inevitable shifts in community data. MCPS resilience to community shifts is challenging due to the lack of large labeled datasets. One promising approach for addressing the lack of large labeled datasets is through the use of optimal transport algorithms for transductive label augmentation. Realizing this promise faces important technical issues regarding privacy, data imbalance, and distributed datasets. Recent research on the role of OT in federated learning may provide an avenue for addressing these technical issues.

3 Relation to Prior Surveys

Resilience in cyber-physical systems (CPSs) has emerged as a critical area of research, driven by the increasing reliance on CPS infrastructures in safety-critical domains such as energy, transportation, and healthcare. In response, several surveys have reviewed techniques and frameworks for achieving resilience in CPS. However, many of these works adopt narrowly scoped perspectives. As a result, they often fail to capture the broader, system-level dimensions of resilience that are essential for real-world deployments. Key aspects such as human-system interaction, data scarcity, and operational adaptability remain underexplored in much of the existing literature. Segovia-Ferreira *et al.*, [198] conduct a

focused study on control-theoretic methods and industrial control systems. Their work provides a structured taxonomy of resilient control techniques and cyber-resilience techniques. It also discusses various evaluation metrics. However, it pays limited attention to resilience challenges that arise from human-system interactions and data generation challenges. Ratasich *et al.*, [190] examine resilience in IoT-based CPSs with a particular emphasis on fault detection, localization, and recovery in resource-constrained and real-time environments. Their survey addresses key challenges such as limited computation, strict latency requirements, and constrained communication, and illustrates their discussion with application case studies like resilient smart mobility. While this perspective is valuable within the IoT domain, it remains narrow in scope with respect to the broader CPS settings. The work does not extend to broader CPS settings involving heterogeneous components, nor does it consider system-wide concerns such as formal verification, adaptive recovery beyond predefined fault models, and learning-based approaches under uncertainty. Cassottana *et al.*, [33] take a quantitative approach by reviewing mathematical models and metrics for assessing CPS resilience. Their work offers a systematic framework to evaluate the resilience of a CPS, before and after a disruption. However, the survey lacks discussion on data-driven learning and the integration of human factors and cyber-physical system components that together shape resilience. Kim *et al.*, [96]provide a comprehensive survey of network-layer security threats in CPSs. While they discuss resilience in the context of detection and mitigation techniques, the scope of resilience is limited to maintaining secure communication and preventing adversarial disruptions. Similarly, Yu *et al.*, [254] provide a structured overview of attack and defense methods in CPSs, yet their focus remains on security threats and detection techniques. Broader system-level concerns are not within the scope of their discussion.

To address these gaps, our survey provides a comprehensive view of CPS resilience by organizing the discussion around five interconnected themes: (1) system-wide resilience across components, (2) CPS learning challenged by scarce, imbalanced, and multimodal data, (3) CPS resilience via specification, coverage, and redundancy, (4) recovery for CPS resilience, and (5) human in CPS. To the best of our knowledge, no prior work offers a comprehensive and integrative view of resilience that addresses the unique constraints and real-world complexities of CPSs.

4 Theme 1: Resilience as System-wide Property

Resilience is fundamentally a system-wide property that reflects the ability of CPS to anticipate, adapt to, and recover from adverse conditions or disruptions such as deliberate attacks, accidental faults, human errors, or naturally occurring threats [40, 47, 78].

4.1 Systems-Theoretic Hazard Modeling and Analysis

Systems and control-theoretic approaches to CPS model accidents and safety hazards as the result of context-dependent constraint violations across multiple levels of hierarchical CPS control structure [113, 114] (see Figure 2). These levels include humans and autonomous controller actions, cyber and physical system components, and the interactions among them. Accidental faults, malicious attacks, and unintended human errors targeting CPS can originate from different system components, including software, hardware, datasets and models for ML, network and interface devices, sensors, actuators, or other interconnected controllers and devices and get propagated through the cyber layer, corrupt cyber states, and cause erroneous controller actions. Then erroneous control actions when occurred under specific system contexts, defined by the combination of operational, environmental, and physical conditions, could lead to safety and security violations and result in accidents.

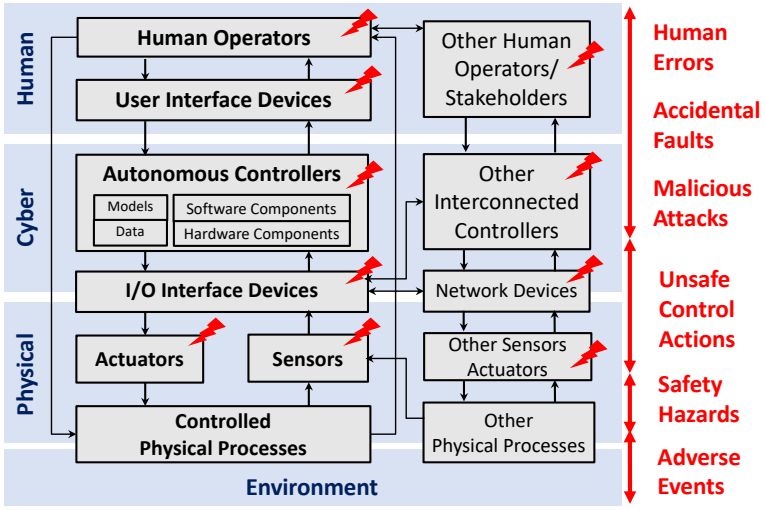

Fig. 2: Hierarchical System Control Structures in Interconnected CPS: Accidental and Malicious Threats to System Resilience across the Human, Cyber, Physical and Environment layers

4.2 Assume-guarantee Contracts for System Resilience

Since resilience is a system-wide property, which needs to hold between levels of abstraction, between system components, and across temporal modes, a common structure underlying design-for-resilience takes the form of an assume-guarantee contract (AGC) [48–51]. An AGC consists of two parts. Firstly, there is an *assumption* (which may be empty) that captures the input to the exchange under description, such as the output from a triggering component or action, the environmental conditions that must hold, or other system state parameters. Secondly, there is a *guarantee* that describes the outputs of exchange under description, which are guaranteed to hold due to our compositional verification efforts. For example, a software subroutine might have an associated AGC stating the enabling parameters for it to be executed and the guaranteed result of executing the subroutine. Similarly, an AGC might describe the interaction between hardware and software components, such that a software component can trigger a hardware action, like moving a vehicle, or that a hardware component can trigger a software action, like a sensor value resulting in a change of plan. AGCs are organized hierarchically and represent a key building block for resilient CPS: at the highest level, we have an assumption that if the CPS becomes operational, that it will behave resiliently in that it upholds certain guarantees for operation until mission termination. Such a design is scalable and represents best-practices for resilient CPS design. For an example case study, in the realm of connected and autonomous transportation systems, the NASA Lunar Gateway Vehicle System Manager uses AGCs to frame system requirements (in English as well as in mathematical logic), conduct design-time formal verification and simulation, and create runtime monitors to support resilient behavior during system operation.

4.3 System-Wide Recovery

One way to ensure resilience in CPS is to ensure that the system remains *safe* in the presence of failures and attacks. Hence, the system takes *proactive measures* — whether or not a fault or attack occurs. This, coupled with analyses that identify *when* to carry out the proactive actions, can significantly increase the safety of the system. To further

improve the resilience of the system, we should ensure that the *(a)* analyses methods and the *(b)* proactive measures are *tamper-proof* by using hardware and software roots of trust.

One of the best proactive measures is the use of *system-wide resets* [2, 3, 89] that are triggered independent of the existence/detection of faults or attacks. Post reset, the system can be restored to a known safe/secure state — by loading the system with trusted software, perhaps from a read-only memory (ROM) or a hardware root of trust. This ensures that the system is reset to a known safe state and the attackers/faults have been cleared out of the system. Within this broad approach, there are important questions that need to answered:

(1) **When to reset?** As mentioned earlier, the system should reset independent of the existence or detection of faults or attacks. If we wait for the attacker or fault to be detected, it may be too late. Thus, if we reset the system too infrequently, it may lead to a significant increase in the risk of attacks and faults. On the other hand, if we reset the system too often, it may lead to unnecessary downtime and a significant decrease in the quality of service.
(2) **how to reset?** The reset mechanism should be tamper-proof so that attackers or faults cannot interfere with the process or even prevent the reset from happening.

For the first issue, we can use the fact that CPS are often deterministic in nature and are dealing with physical systems that are subject to physical laws. Hence, we can use/develop analyses techniques such as real-time reachability analysis [2] that can predict when a CPS is likely to get close to violating is safety constraints [1]. Once we know how long it takes for the system to reach the safety boundaries, we can compute the time when the system should reset — essentially *before* it breaches the safety boundaries and with enough time left to reset the system, load the trusted software and push the system back into a safer state. These specifics of these processes will vary depending on the type of CPS but we have shown it to work for a wide variety of systems — from avionices and building automation systems [2] to IoT systems [3] and even distributed cryptography systems [89].

To decide on how to reset, again there are multiple options, depending on what has been established as a root of trust in the system. For instance, *hardware mechanisms* such as external timers and watchdogs can be used to trigger resets in a tamper-proof manner [2]. Alternatively, specialized hardware features such as ARM Trustzone [175] can also be used to initiate resets [3]. If the operating system (OS) is considered to be trusted then it can also carry out this functionality. In a ditributed setting, one of more of the nodes can be designated as a trusted node and it can initiate the reset process [89].

4.4 Resilience through Obfuscation

Another proactive way to improve the resilience of CPS as a whole is by *obfuscating* the *behavior* of the system — for instance, by adding *noise* to the execution patterns and timing of the system. Obviously, indiscriminate addition of noise can be quite detrimental since that can severely degrade the performance and timing guarantees of the system. Hence, we need to do this in a deliberate manner so that the system appears to be *normal* to an external observer/attacker but, in reality, it is not. This is where the notion of "*schedule indistinguishability*" [39] comes into play. We adapt ideas from the differential privacy world and add noise, in an algorithmically determined manner, to the execution schedules of tasks in the system so that an attacker/observer cannot distinguish between the real execution schedule and the obfuscated one. This method has been shown to be effective against timing-based attacks in autonomous and other CPS, including streaming across the internet.

[1] Note that we assume the main aim of an adversary is to push the system beyond its safety limits. Hence, the prevention of safety violations will also prevent many classes of attacks from succeeding.

5 Theme 2: Data for Learning-Enabled CPS

Overview. When CPS operate in dynamic and unseen environments, their sensory data may be noisy, scarce, or distributionally shifted. In addition, the collected data sometimes is irregularly sampled due to variations in sensor types. Under such scenarios, it is hard to ensure the safety, reliability, and generalizability of learning-enabled CPS. Resilience of learning-enabled CPS refers to systems that can handle uncertainty, identify anomalies, and adapt to unseen circumstances, even where data is noisy, scarce, or out-of-distribution (see Figure 3). The goal of this theme is to present the significant challenges and existing solutions for learning-enabled CPS operating in such environments.

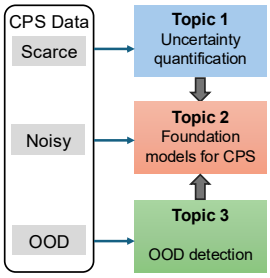

Fig. 3: The overview of learning-enabled CPS.

5.1 Learning in Data-Scarce Environments

Across a range of critical domains, including the two exemplar domains of medical CPS and connected and autonomous transportation, a crucial need is to develop autonomy in changing, possibly hostile conditions where there no capability to learn, train, or model the environment prior to mission commencement. This challenge meets three questions:

(i) *What to learn?* In order to quantify the usefulness of agent's actions — including motion, sensing, and communication — at reducing the uncertainty about its operating environment, we need to introduce a metric of *mission-oriented value of information* [64, 182, 200]. This metric must naturally depend on the variance in mission success across all possible ground truths consistent with available data, as well as the expected decrease in this variance after a single additional observation.

(ii) *How to learn?* Given a set of observations, the *learning* element of a standard learn-and-control loop requires an agent to produce – implicitly or explicitly – an environment model that best fits the given observations. Many current learning approaches rely on function approximation methods [6, 67, 158], with samples provided at each time step during agent's mission, possibly through multiple episodes. However, in a setting of sparse data, the only available knowledge might be outdated or otherwise deficient [56, 160, 181, 253]. Recent semi-supervised learning methods also show that, even under sparse supervision, learning can be improved by leveraging unlabeled data to regularize intermediate feature representations rather than relying only on output-level pseudo-label consistency. For example, SWSEG optimizes both alignment and uniformity of backbone features via a Sliced-Wasserstein objective, yielding stronger segmentation performance in low-label regimes [123]. Thus, research needs to establish meaningful approaches for *varying importance* of available samples, ensuring that available data is maximally used, but not afforded undue trust.

(iii) *How to act?* The agent's decisions are a consequence of two goals. In the short term, the agent seeks to collect data that enable it to understand the ever-changing environment. In the long term, the agent uses this understanding to move towards completing its mission. Such a mission may be complex and not described in terms of time-invariant rewards, but possibly expressible through the formal framework of temporal logic, including both objectives to reach and safety constraints. A natural approach to designing an efficient policy that completes the agent's mission is one of *exploration vs. exploitation* [159, 214]. Namely, after converting the agent's task into a time-varying reward function — equivalently, a time-invariant reward function on a product space of the agent's state space and an automaton that describes the task — it makes sense to provide *bonus rewards* for actions which reduce the agent's uncertainty about the system, using the value of information described above. When

the uncertainty is large, the agent will thus choose actions that reduce it; on the other hand, when the agent has a good understanding of the environment, it will chose actions that drive it towards mission completion. Unlike standard methods of planning in learned environments which make no use of uncertainty quantification and rather often assume perfect learning from a perfect model, it is critical to *certifiably bound* the unavoidable suboptimality of the agent's plans [17, 162, 199]. Consequently, combining the uncertainty-based performance estimation with the active learning and planning may produce *explicit overall bounds on the agent's performance in a data-sparse environment.*

Current state-of-the-art learning methods for control synthesis are often meaningless in data-sparse settings. There is no opportunity to train prior to the mission execution, and each sample collected during the mission might be collected from a different underlying distribution [183, 215, 234, 264]. Understanding that the planner may not be able to exactly learn the correct environment from the small amount of available data, it makes sense to instead identify and exploit invariant features of the environment across space and time to propose notions of (i) mission-relevant uncertainty, quantifying the variance in mission performance across all possible environments supported by obtained data and (ii) sensitivity to data sparsity, describing the trade-off between mission performance and increased need for data. In turn, these metrics are crucial in developing uncertainty-aware plans which enable agents to proceed with their objective while continually learning about the operating environment.

5.2 Resilience to Uncertainties

Quantifying prediction uncertainty is fundamental to building resilient learning-enabled CPS. Two primary types of uncertainty are commonly considered: aleatory uncertainty and epistemic uncertainty. Aleatory uncertainty, or statistical uncertainty, refers to the inherent randomness or variability in the system. It is generally considered irreducible and arises from intrinsic stochasticity in the environment or sensor noise. For example, random fluctuations in vehicle sensor readings caused by electronic interference. In contrast, epistemic uncertainty, also known as model or knowledge uncertainty, stems from incomplete observations, limited training data, or inadequate models. Unlike aleatory uncertainty, epistemic uncertainty is reducible through additional data collection, model refinement, or improved domain understanding. Examples include uncertainty about a patient's physiological state due to missing sensor inputs, or about potential emergent pathologies stemming from limited knowledge of disease progression.

When data is sparse, one approach is to modeling epistemic uncertainty by determining *what possible environments are supported by the collected data and prior knowledge*. One option is thus to extend the commonly used notion of *maximum likelihood*, to inference across unknown stochastic processes [23, 38, 85, 180]. However, while classical work on maximum likelihood only identifies the model that generates the highest probability of observed data, recognizing that determining a single exact environment model is unlikely, it makes sense to generate a distribution of likelihoods over a set of all possible environmental models. Having defined a likelihood-based probability distribution, one possible path of integrating mission-awareness is through the notion of *regret* [10, 65]; namely, considering the sensitivity of performance of a given control law to model mismatch across likely models. Equipped with the mission-aware uncertainty metric, a natural strategy [154, 171, 213] to combine learning and planning is to mix exploration — enticing the agent to take actions which result in high value of mission-aware information — with exploitation — driving the agent to take actions which seem to progress towards mission completion given the current environment model.

5.3 Adapting Pre-trained Foundation Models for Learning-enabled CPS

Foundation models, such as multimodal large language models (MLLMs) [236] and vision-language-action (VLA) models [271], are capable of integrating signals across multiple modalities and controlling physical systems through those multimodal inputs. A key advantage of foundation models lies in their ability to generalize to downstream CPS tasks with limited data, often via zero-shot or few-shot learning [29, 229]. Further, recent advances have introduced lightweight variants deployable on edge devices [248], as well as models capable of incorporating domain-specific knowledge [99, 100]. As such, these foundation models hold great promise for enabling a wide range of learning-enabled CPS applications.

Adapting pre-trained foundation models for CPS applications. A promising direction is to adapt existing, pre-trained foundation models to CPS applications. However, beyond the need for computational efficiency, this adaptation, presents two major challenges. *First*, CPS operating in dynamic environments often encounters noisy, sparse, and out-of-distribution (OOD) data. Such discrepancy between training and inference data can significantly degrade model performance and reliability. *Second*, sensor characteristics during CPS deployment may differ from those used during pre-training in terms of sensing range, resolution, sampling rate, or even modality. This mismatch introduces a fundamental domain gap that can prevent foundation models from fully leveraging available data. Addressing these challenges requires both theoretical analysis to characterize the conditions under which such adaptation is feasible, and practical algorithms that enable effective and robust adaptation in real-world environments.

Learning with noisy, sparse, and out-of-distribution data presents a long-standing challenge in machine learning [227]. Prior approaches for adapting foundation models with limited data include learning from examples in the context prompt (in-context learning) [29], constructing simple classifiers based on the pretrained representation [31], learning lightweight adapters [80], or prepending learnable input tokens (*e.g.*, prompts) [82]. An emerging solution, related to meta learning [79], involves finetuning a pretrained model on multiple auxiliary tasks pertaining to the target task [148, 197]. Our prior work investigated the theoretical justification of this multitask finetuning [241]. In this setting, a pre-trained model is further fine-tuned with a set of relevant tasks before adapting to a target task. Each of these auxiliary tasks might have a small number of labeled samples, and categories of these samples might not overlap with those on the target task. Our analysis builds on a key intuition is that a sufficiently diverse set of relevant tasks can capture similar latent characteristics as the target task, thereby producing meaningful representation and reducing errors in the target task. Our theorem, further confirmed by empirical results, reveals that with limited labeled data from diverse tasks, finetuning can improve the prediction performance on a downstream task. Despite these advances, adapting pre-trained foundation models with limited data remains an open challenge.

Adapting pre-trained foundation models to new tasks involving different sensor data types (e.g., varying resolution) or incorporating additional sensing modalities presents an open problem. A promising direction is adapter-based methods, such as low-rank adaptation (LoRA) [80], which introduce parameter-efficient modules into a pre-trained model without modifying its architecture or core weights. Each adapter can be viewed as a functional "patch" to the foundation model, conceptually analogous to a software update. Such adapters have been widely used to customize large pre-trained language and image generation models, enabling flexible control over content and style without retraining the entire model and only using a small amount of data [80, 104]. Our recent work [122] demonstrated an early version of this idea in the context of video foundation models, introducing lightweight adapters for downstream tasks with additional input modalities, including audio, 3D cues, multi-view inputs. By adding only a small number of parameters and operations to the base model, and leaving its original architecture and weights untouched, the adapted model

can support a range of complex tasks such as audio-visual question answering, 3D reasoning, and multi-view video recognition, with significant performance gains. While adapter-based methods can bridge modest modality mismatches, more substantial sensing heterogeneity may require domain-specific architectural design.

Effective adaptation and efficient deployment of pre-trained foundation models for CPS remains an open research challenge, pointing towards several promising avenues for future exploration. One compelling direction is to advance the theory and practice of data-efficient adaptation: can models be adapted to handle noisy and OOD data using as few samples as possible? Additionally, can synthetic data be leveraged to support this adaptation and reduce reliance on real-world data? Another important direction lies in compositionality: can multiple task-specific adapters be integrated into a unified model to enable improved generalization across a broad range of compositional tasks? Progress in these areas could significantly enhance the applicability and robustness of foundation models in learning-enabled CPS.

5.4 Building customized foundation models for CPS applications

While adapting existing pre-trained foundation models allows leveraging investments in prior model training, it is not always the best approach. Vanilla foundation models are not specialized. They embody large amounts of general knowledge that contributes to model bloat without necessarily being directly applicable to the domain at hand. An alternative solution is therefore to develop domain-specific foundation models from scratch that act as specialists in the given (CPS) application domain. Such models, being limited to a narrower domain, would be smaller and more efficient, sometimes called *micro foundation models* [99, 100]. Developing foundation models for CPS applications introduces a slew of additional challenges [20]. Examples of specialized domains include climate and sustainability data analysis applications, energy optimization in data centers, network security analysis, and maintenance/diagnostics in complex systems, among others. Specialized foundation models can help recognize complex patterns in domain-specific time-series data to advanced diagnostic, reasoning, and prediction capabilities.

Models for CPS applications should be able to efficiently handle: (i) multimodal data of arbitrary modalities (such as data collected from domain specific sensors), (ii) time-series data with frequency domain signatures that encode complex recurrence patterns, (iii) spatial reasoning from multi-vantage data (such as data from multiple observation points), and (iv) large sensor and environment configuration space, as is common in CPS applications. While some of these challenges also arise when adapting pre-trained models, developing models from scratch requires natively supporting these capabilities. This introduces new research questions in model design, training, and evaluation, especially under the typical constraints of limited labeled data and diverse deployment scenarios common in CPS applications.

The state of the art must be advanced in several respects. First, much of today's work on foundation models focuses on data modalities common to human communication and perception, such as vision, text, and audio. The diversity of envisioned intelligent CPS applications may employ, possibly proprietary, multimodal sensor data where sources are substantially more heterogeneous, calling for learning and inference mechanisms suitable for arbitrary multimodal time-series data. Also, many applications, such as those featuring environmental data, residential energy consumption data, data center temperature measurements, and similar sensor-based sources, will feature measurements of some external physical environment. Often the underlying physics have clear signatures in the frequency domain, making Fourier transforms (or spectrograms) a preferred input type. Pre-training must be cognizant of properties of such transforms in order to make learning more efficient. For example, the design of data augmentation pipelines (e.g., in such training components as contrastive learning) must be physically-informed to offer additional structure that can be leveraged in the augmentation design. Additionally, such applications will often use data sources at multiple locations. For example, weather data features deployments of large numbers of sensors (observing their physical environment

from multiple vantage points). Thus, foundation models must be developed that can reason spatially about multi-vantage data sources to compute holistic representations of the observed physical phenomena from the multitude of individual multimodal and multi-vantage data streams. The FMs must support heterogeneity not only in sensing modalities but also in individual sensor properties such as calibration, measurement resolution, and sampling rates, and should neither require per-device re-training nor cause parameter explosion due to the large source (e.g., sensing device) configuration space. Moreover, the mix of sources used may differ substantially from installation to installation (e.g., from one data center to another in data center diagnostic applications) and may evolve over time. The foundation models should be easily customizable to their deployment environment. They should easily support upgrades and remain robustly operational in the presence of individual device/source failures, processing workflow changes, and disconnections.

While normally one might rely on AI scaling laws to endow models with better capabilities, in the case of domain-specific CPS models, training data are usually a bottleneck. Thus, rather than relying on model size (and thus larger amounts of training data) one needs to rethink the training pipeline to learn more efficiently from limited amounts of data by injecting inductive biases inspired by the application domain. Recent work on domain-specific foundation models [99, 100] has espoused this approach to help address the above challenges, improving the efficacy of multimodal self-supervised pre-training [87, 88, 120] and domain-specific data augmentation [118, 136, 222–224, 247] to increase both training performance and data size.

5.5 Resilience to Out-of-distribution (OOD) Detection

OOD detection [91, 145, 188] is critical for ensuring the reliability and resilience of learning-enabled cyber-physical systems (CPS). In general, existing research on OOD detection in CPS can be categorized into two types: single-modal and multi-modal OOD detection.

(i) *Single-modal OOD detection.* This line of work primarily focuses on detecting OOD samples using data from a single sensor modality, such as time-series, image, or LiDAR data. For instance, Ramakrishna et al. [188] adopted a β-VAE to learn disentangled representations from images for detecting OOD frames. More recently, Kaur et al. [91] proposed a conformal anomaly detection framework for OOD detection in CPS time-series data, based on deviations from in-distribution temporal equivariance. Despite significant achievements [91, 188], single-modal approaches have been found to be susceptible to missed detections when that modality is affected by environmental factors such as fog or rain.

(ii) *Multi-modal OOD detection.* Several studies [58, 208, 221] have leveraged multi-modal data to enhance OOD detection accuracy in CPS equipped with multiple sensors. For example, Wang *et al.*, [221] developed a multi-modal transformer network that fuses radar and LiDAR data to detect radar ghost targets in autonomous driving. Qiu *et al.*, [185] introduced an unsupervised contrastive learning method that employs generative adversarial networks to extract latent features from five different sensor modalities for accurate abnormal driving segment detection. While these approaches achieve strong performance in OOD detection, they generally lack interpretability, making it difficult to explain prediction results. To address this limitation, recent works [257, 258] have incorporated LLMs to enhance the identification and reasoning of OOD samples. For instance, Sinha *et al.*, [205] proposed a hybrid framework that combines a fast VAE-based detector with a slow LLM-based reasoner for real-time anomaly detection and reasoning in CPS. Despite their significant strides, LLM-based OOD detection still faces key challenges. First, the system must determine when to trigger the slower reasoning module in place of the faster detector. Second, the latency introduced during the reasoning process remains a critical concern for real-time CPS applications. Third, it still remains a challenge to reduce the false positives or false negatives for OOD detection.

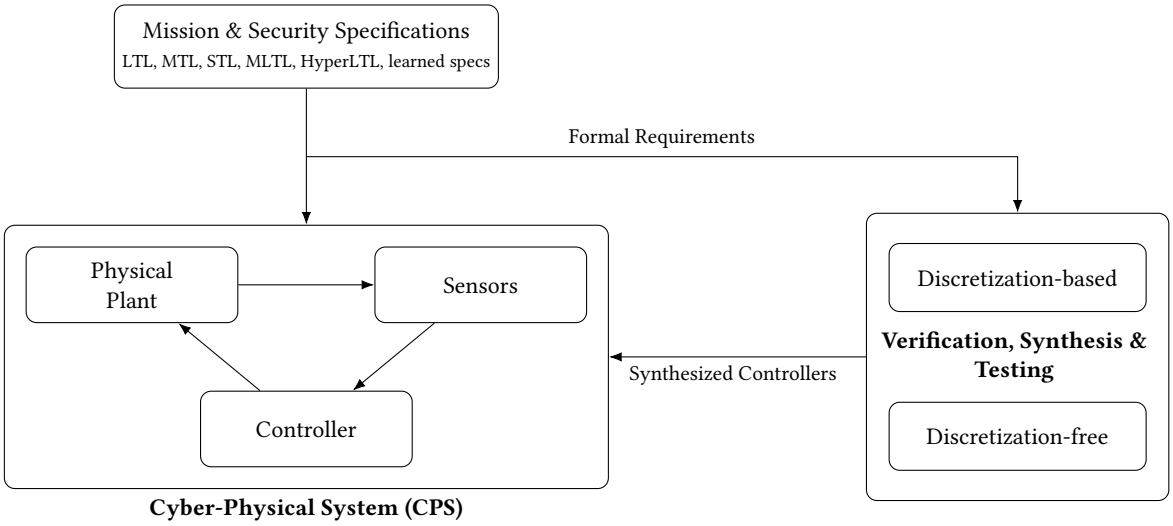

Fig. 4: Theme 3 overview: formal specifications drive verification, synthesis, and testing, which in turn yield correct-by-construction controllers for the CPS.

6 Theme 3: Verification, Testing, and Redundancy for CPS Resilience

Ensuring the resilience of safety-critical CPS over their lifecycle requires the ability to specify and analyze requirements that capture reliable, high-confidence, provable behaviors. We therefore use temporal logics to express requirements in a precise, unambiguous form that enables automated analysis, including formal verification and synthesis.

6.1 Specifications of Resilience, Mission, and Security Properties

Mission Requirements: Without clear mission and resilience requirements, it is difficult, if not impossible, to determine what resilience mechanisms must be designed. Temporal logics have proven effective for specifying behaviors precisely and unambiguously, and can be extended to address resilience challenges.

Linear Temporal Logic (LTL) [19] is a convenient and expressive formalism for specifying properties over infinite executions (traces) of a system; a variant for finite traces has also been proposed [53]. The set of LTL formulas over a finite set of atomic propositions $\mathcal{AP}$ is defined by the grammar:

$$\phi ::= a \in \mathcal{AP} \mid \neg\phi \mid \phi \vee \phi \mid \mathsf{X}\,\phi \mid \phi\,\mathsf{U}\,\phi.$$

Here, $\neg$ and $\vee$ denote logical negation and disjunction, while X and U are temporal operators representing `next` (in the next discrete step) and `until` (the left property holds until the right becomes true), respectively. For convenience, additional logical and temporal operators are derived from this core syntax:

$$\texttt{true} \triangleq a \vee \neg a, \quad \texttt{false} \triangleq \neg\texttt{true}, \quad \varphi \wedge \psi \triangleq \neg(\neg\varphi \vee \neg\psi), \quad \varphi \rightarrow \psi \triangleq \neg\varphi \vee \psi, \quad \mathsf{F}\,\varphi \triangleq \texttt{false}\,\mathsf{U}\,\varphi, \quad \mathsf{G}\,\varphi \triangleq \neg\,\mathsf{F}\,\neg\varphi.$$

Here, $\wedge$ and $\rightarrow$ denote conjunction and implication, while F and G represent the temporal operators `eventually` (at some point in the future) and `always` (at all times), respectively.

The semantics of LTL are defined inductively; see [19] for details. LTL enables rigorous specification of system behaviors. For example, a safety property can be expressed as $\mathsf{G}\,\neg\phi$, asserting that a bad condition ϕ never occurs, while

a reachability property $\mathsf{F}\,\phi$ requires that a desirable condition ϕ eventually holds. For an infinite trace $r \in (2^{\mathcal{AP}})^{\omega}$, we write $r \models \varphi$ if r satisfies the LTL formula φ. The set of infinite traces satisfying an LTL formula can be accepted by either a non-deterministic Büchi automaton or a deterministic Rabin automaton [19].

Beyond LTL, several temporal logics refine expressiveness for CPS. Metric Temporal Logic (MTL) [13] augments temporal operators with time intervals and can be given continuous or pointwise semantics over finite or infinite runs. A variety of interval types is possible (open/closed, bounded/unbounded, integer/real endpoints); see [161] for a survey. Signal Temporal Logic (STL) further extends MTL with real-valued predicates over real-time intervals [57, 137], yielding a highly expressive language for detailed properties of real-time CPS.

However, greater expressiveness often comes at the cost of analyzability: satisfiability and model checking for general MTL are undecidable, and STL faces similar complexity challenges. In practice, one typically restricts to subsets of STL or MTL for real-time checks. Mission-time Linear Temporal Logic (MLTL), developed by NASA as a CPS specification logic [191], strikes a practical balance between expressiveness and computational complexity, enabling tractable formal verification of complex temporal behaviors. Aerospace operational concepts often describe requirements using integer-labeled timelines (e.g., clock ticks, radar sweeps, or Hertz) as in NASA's Automated Airspace Concept [60], the U.S. Navy's Aircraft Carrier Deck Scheduler [195], and the JAXA–NASA GPM Observatory [55]. MLTL efficiently captures such requirements, supports formal analysis, remains readable for certification authorities [63], and is efficiently monitorable in real time [115].

Security Requirements: LTL, MTL, STL, and MLTL formulas are typically used to specify safety and functional correctness requirements, and they evaluate whether a *single* infinite trace satisfies a property. Security properties, by contrast, often depend on relationships among multiple traces of the system.

Alur *et al.*, [11] showed that the modal μ-calculus is insufficient to capture all opacity policies. To address such limitations, Clarkson and Schneider [45] introduced *hyperproperties*, specified using second-order logic over *sets of execution traces*. Hyperproperties generalize linear-time properties [19] by moving from sets of runs to *sets of sets of runs*.

HyperLTL extends LTL with trace quantifiers, enabling simultaneous reasoning about multiple traces:

$$\psi ::= \exists \pi.\, \psi \mid \forall \pi.\, \psi \mid \phi, \quad \phi ::= a_\pi \mid \neg\phi \mid \phi \vee \phi \mid \mathsf{X}\,\phi \mid \phi \,\mathsf{U}\, \phi.$$

The quantifiers $\exists\pi$ and $\forall\pi$ denote "there exists a trace π" and "for all traces π", respectively. Formulas ϕ retain the usual LTL structure, except that atomic propositions are annotated with trace variables: for each $a \in \mathcal{AP}$ and trace variable π, a_π refers to the valuation of a along trace π. A HyperLTL formula with no free trace variables is called *closed*.

HyperLTL can express a wide range of security properties, including opacity. For example, the following formula expresses language-based opacity, assuming that LTL formulas ς and φ describe sets of traces that reveal and do not reveal secrets, respectively:

$$\forall \pi\, \exists \pi'.\, L(\pi) \models \varsigma \rightarrow (H(\pi) = H(\pi') \wedge L(\pi') \models \varphi).$$

This states that for every trace π satisfying the secret-revealing condition ς, there exists a trace π' that is observationally indistinguishable from π (i.e., $H(\pi) = H(\pi')$) but satisfies the non-secret condition φ, thereby preserving opacity. These mission and security logics provide the formal backbone for the verification and testing techniques described in this theme.

6.2 Learning Mission and Security Requirements

Robustness and resilience in CPS must ultimately be defined in terms of concrete mathematical constraints, but specifying such constraints is challenging for typical end users. Although LTL formulas or state-space safe sets [43] can formally encode high- and low-level behavior requirements, they are difficult for untrained users to construct. This is problematic for CPS domains—from home robotics to urban air mobility—where non-experts interact directly with the system. Incorrect specifications can cause verification procedures to confirm the wrong properties, leading to violations of the intended behavior and potentially introducing adversarial vulnerabilities. This motivates rigorous specification checking and principled uncertainty quantification.

Recent work addresses this challenge by inferring specifications from natural language commands [165] and from human demonstration trajectory data [43], which are much easier for non-expert users to provide than formal constraints. These approaches use large language models to translate natural language into symbolic constraints in LTL, and exploit the Karush–Kuhn–Tucker optimality conditions of the demonstrator's control problem to identify specifications from limited human data.

Another direction is to learn *neuro-symbolic* specifications that combine traditional symbolic variables and operators with neural functions and predicates. This allows one to express difficult-to-formalize concepts, such as the presence of a pedestrian in an image. For instance, the Neuro-Symbolic Assertion Language (NeSaL) [237] specifies properties for image-based verification and approximation, supporting assurance of closed-loop CPS [69]. Neural operators can also learn real-world perturbations to images and enter into properties describing robustness to distribution shifts [235].

Because specification inference is inherently ill-posed—multiple plausible specifications may explain the same data—it is crucial to quantify uncertainty in the learned specifications. Recent methods approximate the set of all constraints consistent with the observed data and then use this set to guide conservative downstream planning and control [42], synthesizing controllers that satisfy all candidate safety constraints.

Accurate specification learning directly supports verification and testing by ensuring that resilience analyses are grounded in the *intended* behavioral and safety requirements. Data-driven and neuro-symbolic specification learning thereby complements downstream verification methods, jointly advancing the provable robustness of CPS.

6.3 Modeling Complex Sensing and Perception Modules

Resilient CPS rely on models at varying levels of granularity to design components, test diverse scenarios, and monitor or predict uncertain environments online. Resilience testing, in particular, requires generating unexpected yet realistic inputs. At the same time, CPS resilience is often undermined by unreliable sensing and perception, so accurate, tractable models of sensing and perception errors are essential.

As sensor data become increasingly rich, it is difficult to construct realistic models against which to design, verify, test, or monitor CPS resilience. This is especially true for vision sensors and neural-network-based perception [143], which operate in high-dimensional pixel spaces lacking realistic symbolic models of pixel evolution. Recent work introduces *verifiable probabilistic abstractions* of sensing and perception, ranging from simple statistical intervals [170], to conservative inflations [46], and contraction methods for downstream resilient control [44]. These are parametric perceptual models learned from data together with confidence intervals.

When combined with models of dynamics and control, abstractions of sensing and perception enable *closed-loop verification* of resilience properties. For example, one can verify that a system accomplishes its task even when visual sensors produce degraded images (e.g., grainy or blurry), by carefully coupling statistical models of perception with

deterministic reachability analysis for neural-network-based CPS [220]. A critical deployment step is to check model validity in the target environment; invalid models can invalidate previously obtained guarantees [173]. Overall, such techniques provide a path to resilience against out-of-distribution inputs, as discussed in the previous theme.

A complementary approach uses generative learning. *World models*, which predict future observations from past observations and actions, originated in reinforcement learning to efficiently sample realistic episodes [74]. Generative world models have been adopted in robotics and CPS for monitoring [4], prediction [139], and verification [90]. A key challenge is to firmly ground these generative models in physical reality [172] to prevent hallucinations and support trustworthy resilience assessment. This would allow the CPS community to bring decades of experience in relating cyber and physical dynamics to the emerging challenge of physically unrealistic hallucinations in generative AI.

6.4 Verification, Synthesis, and Redundancy with Known/Learned Specifications and Models

6.4.1 Discretization-based Techniques: A standard approach to designing correct-by-construction control software or performing formal verification is via *symbolic abstraction*. Here, requirements are specified using temporal logic (e.g., LTL) or automata over infinite strings [19]. A finite, discrete abstraction of the continuous dynamical system is constructed so that any controller synthesized on the abstraction can be systematically refined to a controller for the original system, and verification results transfer from the abstraction to the concrete system. These finite abstractions provide a unified modeling framework for both cyber and physical components. By composing their finite-state models, we obtain a complete finite-state representation of the CPS, enabling the use of techniques from discrete-event systems [32, 103] and automata games [135, 138, 216] for automatic controller synthesis and property verification. The synthesized discrete controller is then refined into a hybrid controller, or the guarantees are carried over to the concrete model.

Research on symbolic abstractions has evolved in three major directions. The first constructs non-deterministic finite abstractions that over-approximate the input-output behaviors of the original system, often motivated by sensing and actuation limitations [126, 187]. Extensions based on Willems' behavioral theory and ℓ-complete systems [147, 232] improve abstraction accuracy by increasing the history length ℓ of input/output symbols, at the cost of larger models. The second, inspired by bisimulation theory [141, 168], builds quotient systems by partitioning the state space to obtain exact bisimulations between the original and abstract models [12, 101, 209]. Such partitions, however, may not exist or may not terminate even for simple linear systems [15]. The third, more recent direction introduces approximate equivalence relations, such as approximate bisimulation [70], which relax exact equivalence by allowing bounded output discrepancies between related states. This greatly expands the class of systems that admit finite abstractions and has led to a rich body of results [71, 177, 192, 209, 255, 256].

Unfortunately, none of the finite abstractions in this literature is guaranteed to preserve security properties such as opacity. As shown in [259], standard (bi)simulation relations and their approximate variants, typically used in abstraction schemes, fail to preserve opacity. Recent results [121, 250] adapt simulation relations to the opacity setting and develop, for the first time, abstraction-based techniques for opacity verification in CPS.

6.4.2 Discretization-free Techniques: The above abstraction-based framework provides a systematic way to address mission and opacity properties in complex CPS, but it can face scalability challenges due to discretization of state and input spaces. This has motivated discretization-free approaches, particularly those based on barrier certificates.

Over the past two decades, *barrier certificates* [178] have become a powerful tool for safety verification of dynamical systems, with automated search enabled by optimization techniques such as sum-of-squares programming [169].

Consider a system $x(t+1) = f(x(t))$ with $x(t) \in X$ for all $t \in \mathbb{N}_0$. A function $B : X \to \mathbb{R}$ is called a *barrier certificate (BC)* if

$$B\big(f(x)\big) \;\leq\; B(x) \quad \text{for all } x \in X.$$

If there exists a BC such that $B(x) \leq 0$ for all $x \in X_0$ (initial set) and $B(x) > 0$ for all $x \in X_1$ (unsafe set), then the system never reaches X_1 from any $x_0 \in X_0$. In automata-theoretic verification, a central challenge is to determine whether a set of states can be visited only finitely many times. This can be reduced to a safety-like condition by bounding the number of visits to a given region. Recent results [150] build on bounded synthesis and introduce an abstraction-free technique for automata-theoretic verification of discrete-time dynamical systems, using barrier-certificate-style constructions for ω-regular properties.

Despite the power of verification, it is typically reactive—analyzing a system after it is designed. Correct-by-construction synthesis instead aims to *generate* controllers or estimators that are guaranteed to satisfy safety or performance specifications by design, embedding correctness into the synthesis process. Safe controller synthesis constructs control strategies that prevent the system from entering unsafe states, often under uncertainty or adversarial disturbances [22, 210]. This proactive perspective is central to resilient CPS.

6.5 Provably Resilient Estimators in CPS

When complete state information is unavailable, estimator synthesis is used to infer system states while respecting high-level requirements. This is particularly relevant for state estimation under sensor or actuator attacks and under bounded or stochastic disturbances [94, 95, 164, 251].

Provably safe state estimation is crucial for CPS integrity in adversarial environments. Formal observers aim to characterize when agents, actuators, or sensors are compromised and to map such conditions into security or privacy policies. The field of secure state estimation has evolved from *reactive* detection and mitigation of attacks [83, 144, 151] to *pre-emptive* resilience in the estimation process, including characterizing upper bounds on the number of attacked sensors/actuators [52, 142, 251]. These bounds, together with the notion of *sensor redundancy* (multiple sensors measuring the same quantity), guide preventive attack mitigation: one can determine which sensors or actuators must be protected to guarantee resilient estimation prior to deployment [95, 242, 243, 251].

Hardware redundancy, such as triple modular redundancy (TMR) in flight and aerospace systems [24, 249], will continue to play a key role in pre-emptive resilience. Determining which hardware elements should be duplicated or triplicated in CPS, while balancing cost and weight, remains an important open problem. Together with the verification, testing, and synthesis techniques above, resilient estimators and hardware redundancy form a critical layer of defense for CPS, enabling robust operation even under faults and attacks.

7 Theme 4: Recovery for CPS Resilience

Overview: Resilience requires the ability to recover from and adapt to changing conditions, failures, and attacks. The complexity and scale of uncertain, high-dimensional, networked CPS, such as mobile sensor and robot networks operating in unknown environments, pose significant challenges to the timely and safe response of CPS, so that they can efficiently tolerate disruptions and adapt to new conditions. For example, slowly time-varying environmental conditions may require replanning of the nominal objectives of the CPS, along with on-the-fly data collection for learning the new conditions, while unanticipated failures of CPS components require rapid detection and mitigation strategies. Furthermore, CPS can be vulnerable to adversarial attacks on their sensor and actuator systems, raising the

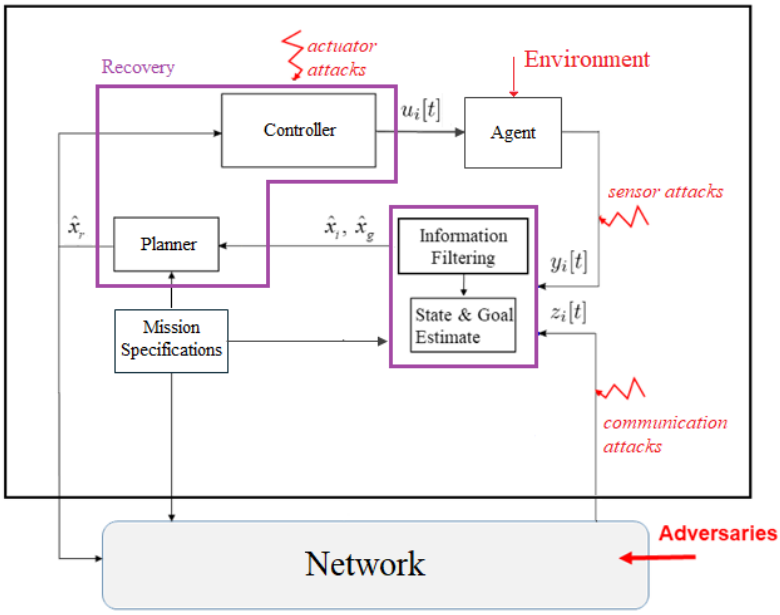

Fig. 5: A recovery architecture against attacks and environmental uncertainty.

need for resilience against such threats. Finally, networked CPS raise additional challenges in how the communication topology and the level of connectivity among agents offers resilience guarantees against the possibly compromised information flowing through the system.

Motivating Example: To articulate the above aspects of CPS recovery in more detail, let us consider a nominal stochastic control process where an agent (or a collection of agents) must make sequential decisions (i.e., over time) under uncertainty. Typically, CPS, under stochastic control processes, operate under a *sense-predict-plan-control* loop, where a perception component *senses* the environment, a prediction component forecasts the evolution of the environment, a planner optimizes the expected performance of the agent under the predicted evolution,[2] and a controller optimizes the low-level control actions. For example, consider a drone that must navigate an uncertain environment and reach a target of interest. In this case, the drone's onboard sensors sense obstacles and terrain features, while a learned or model-based predictor forecasts how wind conditions or moving obstacles, e.g., other drones, may evolve. A high-level planner, typically pre-trained offline, generates a trajectory that balances risk and reward, and a low-level controller executes fine-grained motor commands to follow the trajectory safely and effectively.

7.1 Resilience as (Re)Planning in Stochastic Environments

We focus specifically on the planner for this discussion. The planner's decision-making process can be modeled as a Markov decision process (MDP), where the state captures the current configuration of the drone and relevant features of the environment, the action corresponds to a high-level navigational decision, and the transition dynamics reflect the stochastic evolution of the system due to environmental uncertainty. Given this formulation, the planner's goal is to compute a policy—a mapping from states to actions—that maximizes the expected cumulative reward, typically

[2]The "expectation" is taken with respect to the stochastic plan, which subsumes the stochastic evolution of the environment.

representing progress toward the target while avoiding risk. Such a policy can be computed using dynamic programming techniques when the model is known and tractable, or learned through reinforcement learning when the model is unknown or the environment is too complex to model explicitly.

In many real-world scenarios, the environment in which a CPS operates is not stationary; its underlying dynamics and reward structures may evolve over time [5, 30, 93, 127]. If such time-varying conditions are known beforehand—e.g., during training—and exploration across the temporal axis is feasible (such as through simulation), then the problem can be reduced to a stationary decision-making setting [93, 108]. In theory, the model designer can augment the state space with a time index, effectively treating time as a feature, and learn a time-dependent policy over a stationary MDP. However, in practice, environmental changes are often unexpected and unmodeled, making such augmentation ineffective. In these cases, the CPS must first detect that a change has occurred, then update its internal model of the world—potentially inferring how the environment's dynamics or reward structures have shifted—and subsequently replan or adapt its policy in accordance with the new conditions.

These challenges are naturally captured by the framework of non-stationary Markov decision processes (NS-MDPs), where the transition dynamics or reward functions evolve over time in unknown ways [36, 93, 108]. A central difficulty in such settings is that the agent must collect new data to update its understanding of the world, yet doing so requires acting in an environment whose behavior it does not fully understand. This creates an inherent tension between exploration and safety. A common approach to address this is risk-averse planning [108], in which the agent initially follows a pessimistic policy that prioritizes safety and robustness while gathering data, and subsequently transitions to a more performant policy once sufficient knowledge has been acquired [127, 128]. However, data collection and model updates do not guarantee improved performance. The agent's uncertainty during planning can be broadly categorized as epistemic or aleatoric [128]. Epistemic uncertainty arises from incomplete environmental knowledge and can, in principle, be reduced through active exploration. On the other hand, Aleatoric uncertainty is intrinsic to the environment—for instance, due to sensor noise, wind gusts, or moving obstacles—and cannot be eliminated through data alone [128]. In the drone scenario, epistemic uncertainty may stem from an incomplete map of wind patterns or terrain, which can be improved with data collection. Aleatoric uncertainty may arise from unpredictable gusts or transient obstructions, which persist despite data collection. Thus, while the drone can reduce epistemic uncertainty through targeted exploration, it must continue to operate under a risk-averse policy in regions of high aleatoric uncertainty to ensure safety and resilience.

In such non-stationary settings, the agent is often required to compute a plan online, adapting in real time to observed changes. Monte Carlo Tree Search (MCTS) [102] is a widely used approach in this context, allowing the agent to simulate possible futures and select actions based on sampled trajectories. Importantly, the agent need not discard its previously learned policy altogether. If the environment has only changed partially, the existing policy can still provide valuable priors [174]. This motivates using policy-augmented search techniques, where the prior policy guides the MCTS rollout distribution or initializes the search tree, balancing adaptation with prior knowledge [128]. Such hybrid strategies are particularly effective when changes are localized or gradual, enabling the agent to maintain performance while remaining responsive to environmental shifts.

7.2 Guaranteed Transition between Nominal and Backup Systems

We know that the trading-off between exploration and exploitation in reinforcement learning is an open challenge, especially in non-stationary environments. Nevertheless, it offers a principled manner of how resilience can be viewed as the transition from the nominal system operation (under a nominal policy) to the backup system operation (under a

backup policy). Continuing the example above, guaranteeing that the agent will always remain safe while navigating along a possibly unsafe nominal trajectory (which can be viewed as exploration) requires that the nominal trajectory can be always transitioned in a timely manner to a backup trajectory that is safe, albeit of possibly reduced functionality compared to the nominal one. In other words, for safe navigation among unknown obstacles that are detected on-the-fly, exploration could be viewed as finding an optimal trajectory (nominal) that maximizes the sensed area, albeit possibly risking safety, while exploitation could be viewed as finding a trajectory (backup) that remains in the sensed safe area, albeit possibly being suboptimal w.r.t. the nominal objective. Viewing this paradigm in resilience terms, the backup system is the one of reduced functionality that can keep the system "running" if the nominal system fails. Hence, recovery entails that the nominal system (exploration) should be designed to always be able to transition to the backup system (exploitation) so that upon the disruption event, the system trajectories can be governed by the backup system.

The design of nominal-backup system pairs and policies for recovery raises significant challenges, including how to obtain and interconnect them in a computationally-efficient manner for real-time execution and adaptation. Considering the sensing-planning-control loop, the sensing component builds the sensed area, or currently known set, of the environment, which can have safe and unsafe parts, the planner generates a nominal path or trajectory for the robot to follow to a goal location, which can lie through the currently unknown, and hence possibly unsafe, part of the environment, and the controller tracks this nominal trajectory. Since the known set is built on-the-fly, the nominal trajectory needs to be verified online that it is safe to be tracked by the controller. Nominal system trajectories that are predicted to be unsafe should be mitigated by safe, backup system trajectories.

We have recently developed a computationally-efficient method for online safety verification [7], called Gatekeeper, which serves as a novel component between the planner and controller. Gatekeeper defines the committed trajectory that the system tracks as a composition (and hence trade-off) between the nominal trajectory and the backup trajectory. The construction of the committed trajectory and the verification of its safety is done recursively, and by construction assures safety under certain assumptions. By tracking the committed trajectory as generated by Gatekeeper at each iteration, the system is guaranteed *to be able to, if needed,* converge to a safe backup set. The Gatekeeper principle has already been applied to recovery objectives such as energy renewal in mobile sensor networks [153] (enacting safe transitions to a charging station for a team of autonomous aerial robots), and landmark tracking in aerial surveillance [41] (for reducing navigation/localization errors under budget, state, and input constraints). While co-designing the nominal and backup systems remains an open challenge and problem, and depends highly on the mission at hand, generic methodologies for robust and adaptive safety control that can be used to define the backup set include Control Barrier Functions [68], and their formulations that handle input constraints and disturbances [8, 28], online adaptation to structural parameters [25, 97] and risk-aware settings [26, 225], in tandem with classical planners such as MPC and RRT* for compatibility between nominal and backup trajectories [9, 41, 98].

7.3 Recovery from Adversarial Attacks

Recovery for CPS should be both safe and prompt to ensure system resilience. Recovery safety and recovery promptness are not independent of each other. Recovery safety entails that the recovery process does not introduce new hazards or violate critical system constraints. Recovery promptness is essential to minimize performance degradation and prevent cascading failures.

Adversarial attacks on CPS can be categorized by their purpose. The first type directly compromises safety, such as actuator signal injections that induce structural resonance, potentially causing severe damage. These are the main

focus of recovery methods discussed in this subsection. The second type targets non-safety aspects, including system confidentiality, infrastructure availability, or privacy.

The attack surfaces through which adversarial actions are executed span both the cyber and physical domains. In the cyber domain, attackers may employ techniques such as false data injection, signal delay, or denial-of-service to disrupt communication and control logic. In the physical domain, attacks manipulate the surrounding environment to induce sensor misreadings. For example, through temperature changes, electromagnetic interference, or physical obstructions. Importantly, physical-domain attacks can be particularly challenging to detect because they often circumvent traditional software-based anomaly detection mechanisms by exploiting the physics of the system rather than its code.

Recovery strategies from adversarial attacks can be broadly classified into two categories: shallow recovery and deep recovery [124]. Shallow recovery methods aim to restore system functionality by reconstructing or estimating the corrupted sensor or actuator data and allowing the original control logic to continue operating with minimal disruption [1]. These methods typically rely on statistical filtering, observer-based estimation, or state reconstruction using historical data. While shallow recovery is computationally lightweight, it assumes that the original controller remains effective and efficient after data restoration. This assumption may not hold in the presence of persistent attacks.

In contrast, deep recovery adopts a more comprehensive approach by engaging a dedicated recovery controller designed to steer the system back to a safe operational state [119, 260–262]. These recovery controllers may be synthesized using model-based or data-driven techniques and are often equipped with formal safety or performance guarantees. Deep recovery mechanisms are especially valuable in scenarios where the original controller cannot ensure safe operation due to altered system dynamics, actuator saturation, or unrecoverable state divergence. While deep recovery generally involves higher computational cost and design complexity, it offers greater timeliness and adaptability in adversarial settings.

In addition, there is often a need for *proactive measures* so that the system can recover *before* it moves to an unsafe state — either by an adversary or even a fault. The proactive measures include, *(a)* methods to restore the system to a "clean" state (perhaps by restarting the system and reloading from a read-only memory) [2, 3] and *(b)* addition of noise in a systematic manner [39] to ensure that an adversary cannot predict the behavior of the system, and hence cannot exploit it. These methods focus on ensuring the safety of the system — in the presence of adversarial/faulty behavior — and are complementary to the recovery methods discussed above.

7.4 Recovery of Networks in Networked CPS

Networked CPS, such as mobile sensor networks and teams of networked autonomous robots, rely on the interconnection among components to enable collective decisions and system-level behaviors that emerge from local interactions. Communication topologies and levels of connectivity [179] that govern the evolving inter-component interactions in these networked systems are thus critical for determining the system resilience. Therefore, leveraging system-level redundancy is essential for designing recovery strategies that enhance the resilience of networked CPS by initiating system-wide transformations to preserve or restore networking capabilities.

Designing effective recovery strategies for networked CPS is challenging because it involves balancing competing goals. While network redundancy enhances fault tolerance, maintaining additional connections can lead to conservative behaviors due to limited promixity-based communication capabilities that restrict the system's performance, especially when flexibility is essential. It is critical to jointly consider (i) how to define effective redundant network structures based on measurement of network resilience, (ii) how to control the behaviors of the networked system to satisfy the redundancy-prescribed specifications, and (iii) how to balance redundancy and performance during

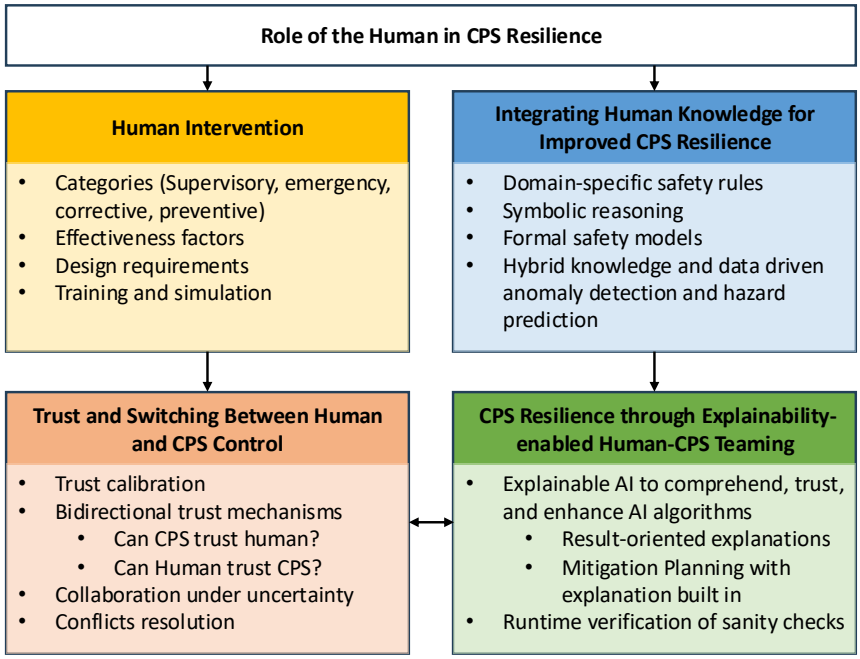

Fig. 6: Overview of Theme 5: Role of the Human in CPS Resilience.

system reconfiguration for recovery that adapts to the rapidly unfolding situations. Graph-theoretic measures such as algebraic connectivity [131, 157, 244] have been widely used to design motion controllers that ensure the integrity of multi-robot communication networks as their topologies evolve during coordinated robot movement. In the presence of a bounded number of random robot failures or malicious robots, resilience measures such as $k-$connectivity [129, 130] and $r-$robustness [107] can be employed to design recovery strategies that re-configure multi-robot networks to attain networking capability with random node removals or mitigate the influence of informational adversaries [34, 109, 218]. The graph robustness notion in particular generalizes the notion of network connectivity, and captures how dense the communication structure should be for the network to be able to tolerate a known upper bound on the number of adversarial agents, i.e., so that the intact agents can achieve consensus on common values despite the effect of malicious information from the adversarial agents [107]. For a recent survey on network resilience the interested reader is referred to [176]. Recent work [245, 246] has proposed uncertainty-aware and data-driven approaches for maintaining multi-robot communication structures that adapt to positional noise and real-world communication performance, while being co-optimized with task-level objectives.

8 Theme 5: Role of the Human in CPS Resilience

While much research has focused on the technical aspects of CPS resilience (e.g., fault tolerance, redundancy), the role of the human remains critical, especially in human-in-the-loop cyber-physical systems (HCPS), such as Advanced Driver Assistance Systems (ADAS) and a variety of medical CPS. This section explores how human interactions influence CPS resilience, categorizes types of human involvement, and identifies key challenges and opportunities for designing resilient CPS. Fig. 6 shows the overview of this section.

8.1 Human Intervention

Human intervention remains a cornerstone of resilience in CPS, particularly in contexts where system autonomy is insufficient to handle anomalies, unexpected events, or degraded conditions. While modern CPS increasingly integrate intelligent automation and fault-tolerant architectures, they cannot fully eliminate the need for timely and informed human interventions. The human operator often serves as a fallback mechanism, stepping in when automated systems reach the boundary of their capabilities, or when situational complexity exceeds preprogrammed responses [114, 203].

Human interventions in CPS can be broadly categorized as *supervisory*, *emergency*, *corrective*, or *preventive*. In *supervisory* roles, humans oversee system behavior and intervene occasionally to correct deviations or refine performance. For example, a grid operator may fine-tune load balancing strategies in response to forecasted weather disturbances. *Emergency* interventions, on the other hand, are characterized by immediate human action during critical failures or hazards, such as when a pilot disables an aircraft's autopilot system during turbulence or a driver takes control of a semi-autonomous vehicle to avoid a collision. *Corrective* interventions occur after faults are detected, often requiring human diagnosis and reconfiguration of the system, such as after a detected cyberattack on an industrial control system. *Preventive* interventions are proactive measures, like scheduling maintenance or isolating vulnerable subsystems prior to a high-risk event.

The effectiveness of human intervention is tightly coupled with several time-dependent factors, including detection latency, cognitive load, decision-making ability, and action execution time. Particularly in time-sensitive CPS, such as autonomous vehicles, smart manufacturing systems, or surgical robots, even brief delays in human reaction can lead to undesirable or dangerous outcomes. In contrast, timely human interventions can significantly mitigate potential safety hazards and improve the system's resilience. Cognitive psychology research shows that under high stress or low awareness, human performance deteriorates significantly, especially when transitioning from a passive monitoring role to an active control role [166]. This switch latency, i.e., the time taken for the human to re-engage with the system and execute an appropriate action, can be exacerbated by poor interface design, ambiguous system feedback, or insufficient training.

Designing CPS that support effective human intervention requires careful attention to system transparency, feedback clarity, and interface usability. The system must present its internal state, decision logic, and level of certainty in a manner that facilitates rapid human comprehension and trust calibration [233]. Furthermore, systems should maintain historical context and provide situational summaries that allow human operators to quickly regain situational awareness during handovers. Training and simulation also play vital roles in improving the quality and timeliness of interventions. Repeated exposure to realistic failure scenarios enhances human readiness and helps develop robust mental models of system behavior. For instance, power grid operators and air traffic controllers regularly undergo simulator-based training to improve their ability to manage rare but high-impact incidents [78]. Such training should not only focus on routine tasks but also emphasize adaptive decision-making under uncertainty.

To this end, developing realistic, domain-specific human intervention simulators is essential for both operator training and system design validation. These simulators must support interactive, real-time scenarios that closely mirror CPS operations under various fault conditions, cyberattacks, or system degradations. For example, in the autonomous driving domain, our prior work [268, 270] introduced a rule-based driver simulator to assess its performance on improving the system resilience against perception attacks. A complementary study developed a human-in-the-loop driver simulation framework to evaluate autonomous driving safety and vulnerability under varied driving conditions and behaviors [155]. Other similar efforts include Shao *et al.*'s CARLA-based closed-loop simulator for evaluating instruction-following and

trustworthy AV behavior [201], Wei *et al.*'s ChatSim for editable, photo-realistic, and human-centric scene simulation [230], and Sim-on-Wheels, which overlays virtual hazards into live driving feeds to create hybrid virtual-real test environments [202]. These tools allow for high-fidelity studies of interface design, human decision latency, and adaptive system response during emergencies. Together, these simulators enhance operator preparedness and also guide the co-design of resilient CPS, providing insights into how human behavior, perception, and trust interact with autonomous decision-making. Future simulators should integrate richer multimodal sensing, real-time explainability, and adaptive scenario generation to evaluate how human-CPS teaming behaves under stress, deception, or degraded awareness.

In conclusion, as CPS continue to evolve toward higher autonomy, human intervention remains indispensable for maintaining system resilience. Ensuring that operators can effectively detect, interpret, and respond to anomalies in complex CPS requires an integrated approach that melds technical robustness, including clear system state and uncertainty reporting, with human-centered design, featuring transparent interfaces and intervention pathways.

8.2 Integrating Human Knowledge for Improved CPS Resilience

Enhancing the resilience of CPS requires more than robust algorithms and fault-tolerant architectures; it demands the explicit integration of human and expert knowledge into both system design and runtime operation. Human intuition, domain-specific safety constraints, and prior operational experience often encode valuable insights that machine learning models or control algorithms alone may not capture. As a result, CPS research increasingly focuses on leveraging structured expert knowledge to guide ML algorithms, enhance interpretability, and improve responsiveness under uncertainty or attacks [186, 269].

A growing body of work aims to formalize the integration of such knowledge into ML pipelines. Integrating logic-based constraints or domain-specific safety rules during training has become a key approach for aligning ML models with human expectations. These include soft-constraint formulations [238], logic circuits for structured generalization [117], embedding logical rules in network architectures [37], and graph-based knowledge representations [73]. Techniques like knowledge distillation have also been used to transfer structured or symbolic knowledge into compact, efficient networks while preserving behavior fidelity [72].

In the context of CPS safety, domain knowledge is particularly valuable when formalized as Signal Temporal Logic (STL) properties. STL provides a rich framework for expressing temporal behaviors and constraints in continuous systems. Several recent studies have explored the integration of STL into learning-based monitoring and anomaly detection. For example, Bartocci *et al.*, [21] proposed methods to mine STL properties from data, while [81] introduced approaches to learn STL formulas from positive examples. These works have demonstrated that logic-based specifications can guide anomaly detection in CPS applications such as industrial control and automotive systems [86].

Our prior work [266] extended this paradigm by proposing a formal framework to synthesize STL-based safety monitors in artificial pancreas systems (APS), based on control-theoretic hazard analysis. Building on this foundation, we have developed a new approach that integrates STL safety requirements as soft constraints into sequential prediction models for real-time hazard detection and mitigation. This extension enables context-aware monitoring that not only detects but also predicts unsafe behaviors, and it is generalizable to a broader class of CPS, including ADAS [267]. Beyond STL, symbolic reasoning and customized learning constraints continue to play an important role. For example, Chen *et al.*, [37] proposed embedding logic rules directly into recurrent neural networks using feedback masking techniques. Similarly, reinforcement learning research has leveraged teacher policies and symbolic constraints to train more efficient and safer agents [194]. These strategies highlight how structured expert knowledge can shape exploration, accelerate learning, and reduce unsafe behaviors during both training and deployment phases.

Ultimately, resilience in CPS depends not only on the system's internal capabilities but also on its ability to incorporate and respond to knowledge from human operators, engineers, and domain experts. By tightly coupling symbolic reasoning, formal safety models, and expert-guided ML, modern CPS can achieve higher assurance, faster recovery from faults, and improved adaptation to new or adversarial environments.

8.3 Trust and Switching Between Human and CPS Control.

Trust is a cornerstone of effective human-CPS collaboration, particularly in high-stakes environments such as autonomous vehicles, healthcare robotics, and industrial control systems. The degree of trust between humans and CPS influences system resilience by shaping how and when control is shared or switched between human and machine agents [112, 203]. Trust governs reliance, influences operator engagement, and ultimately affects the system's ability to recover or adapt in the face of faults or adversarial disruptions. Trust in automation is dynamic and contextual. Users may exhibit overtrust, resulting in uncritical reliance on potentially faulty automation, or undertrust, leading to disuse even when the system is competent [112, 212]. To promote resilience, CPS must calibrate trust, maintaining user confidence while encouraging appropriate skepticism when necessary. This can be achieved through transparent system design, performance feedback, and explainability mechanisms [140].

Recent advances emphasize the dynamic modeling of trust and the development of adaptive trust-aware architectures. Wang *et al.*, [228] proposed trust-aware control strategies that adjust trajectory planning based on real-time trust metrics, thereby reducing operator workload in teleoperation tasks. Similarly, recent frameworks for human-robot collaboration have explored dynamic role allocation, where the system adjusts its behavior, such as switching between autonomous and compliant modes, based on the operator's inferred trust and cognitive state [116, 231]. These mechanisms are particularly vital in collaborative settings like mixed-initiative control of UAVs, where seamless transitions between human-led and autonomous operations are required to maintain safety under uncertainty [226].

A critical enabler of this calibrated trust is explainability, the system's ability to communicate the rationale behind its behaviors. While explainability will be discussed in greater detail in Section 8.4, it is worth noting here that Explainable AI (XAI) techniques are essential for preventing both over-trust and distrust [207]. Quantitative studies demonstrate that interactive explanations significantly improve user confidence and decision accuracy, allowing operators to calibrate their trust levels to the system's actual capabilities [105].

Trust is also bidirectional. Not only must humans trust the CPS, but the system must also assess when it can rely on human input. For example, a distracted driver may not be capable of taking over from an autonomous vehicle in a timely manner. Research in human state estimation and adaptive delegation seeks to address this challenge by monitoring human readiness and context [59, 62]. These bidirectional trust mechanisms are vital for safe and resilient switching between control regimes. Recent studies have begun exploring trust formation in multi-agent and distributed CPS, where teams of humans and AI agents collaborate under uncertainty [61]. The dynamics of inter-agent trust, conflict resolution, and shared situational awareness present new challenges and opportunities for CPS resilience.

8.4 CPS Resilience through Explainability-enabled Human-CPS Teaming

Despite significant technological advancements in CPS, stakeholders remain hesitant to fully accept and rely on these systems for decision-making [54, 217, 265]. One fundamental reason for this reluctance is the lack of explainability, which leads to unknown reliability and trustworthiness. This hesitation is particularly acute when CPS decisions directly impact public safety, urban planning, and critical infrastructure. When a CPS decision results in negative outcomes, city officials and decision-makers must be able to explain and justify how that decision was made. Additionally, many urban

operations are tightly regulated. If CPS cannot demonstrate compliance with regulatory requirements due to opaque decision-making processes, its deployment may not be legally permissible.

Explainable AI is a prerequisite for resilient CPS operations, particularly when human operators must validate autonomous countermeasures against cyber-attacks. Explainable AI research has its roots in the need to comprehend, trust, and enhance AI algorithms [14]. Recent studies have highlighted the importance of assisting human users in trusting decisions made by models like neural networks (e.g. [133, 134]), a critical factor when these models control safety-critical physical infrastructure.

The first category, result-oriented explanations, is vital for post-incident resilience and forensics. Here, the objective is to reveal the rationale behind a decision post hoc. Previous work, such as Langley [106], discusses behaviors encompassing explaining the objectives of a planning task. In a CPS context, this capability allows operators to verify if a system's *resilient* response (e.g., shutting down a valve) was triggered by a legitimate threat or a sensor error. A comprehensive survey by Sreedharan *et al.*, [206] defines considerations for the control operator in resilient systems. Fox *et al.*, [66] identify key questions an explanation system must answer; answering these questions in real-time is essential to reducing the mean time to recovery (MTTR) in compromised systems. The second category, generating explanations as a sub-goal of planning, directly supports human-in-the-loop resilience. For instance, Chakraborti *et al.*, [35] views the planning problem as a system where the AI provides suggestions to the human. This aligns with resilient CPS architectures where AI detects anomalies but requires human authorization to execute high-risk mitigation strategies. Boggess *et al.*, [27] focus on explaining MARL policies; such explanations are necessary to ensure that distributed agents (e.g., in a smart grid) act cohesively during a cyber-physical attack.

Explainability through runtime verification of sanity checks One method of explainability for trustworthiness of CPS is using runtime verification to monitor properties representing *sanity checks*. The efficient reporting of such sanity checks in a human-understandable way enables human or automated intervention. One benefit of this method is that the runtime verification engine can be very rigorously verified and run efficiently, on-board CPS in real time, unlike AI-based solutions. For example, the Realizable Responsive Unobtrusive Unit (R2U2)[84] is a real-time, online, stream-based runtime verification engine that takes as input a set of MLTL specifications to monitor and a description of the CPS hardware resources. R2U2 embeds on-board CPS during operation to continuously monitor that the current mission adheres to the operational requirements, while obeying provable guarantees with respect to the particular CPS hardware. There are currently three implementations of R2U2, in VHDL, C, and embedded Rust; each implements the same core algorithm but tailored to different CPS platforms. R2U2 has successfully verified many real CPS that were specified using MLTL, including NASA's Robonaut2 [92], the JAXA OPS-SAT autonomous satellite [156], a UAS Traffic Management (UTM) system involving Collins and Mosaic Aerospace [75], a sounding rocket [77], and multiple small satellites [18, 132]. We can use tools like R2U2 to monitor for different types of sanity checks, such as that we are not using data from a sensor that is out of range, or whether a control action resulted in the expected temporal reaction; see [193] for a more complete list of sanity checks. Reporting the result of sanity checks in real time can enable more confident intervention. R2U2 has GUI interfaces to enable easy human understandability of its outputs [84].

9 Looking Ahead: Five Mid-term and Long-term Goals

Here, we outline five of the most important open challenges that the community must address to enable widespread adoption of CPS in critical application domains. These challenges organically arise from our discussion in the article on the five themes for resilient CPS. For each challenge, we lay out some mid-term goals, defined roughly as 2–3 years out, and some long-term goals, defined as 5+ years out.

1. Compositional and Adaptive Foundation Models for Edge CPS

Learning-enabled CPS relies on making CPS applications runnable on the edge, which will comprise primarily resource-constrained devices. Current trends involve adapting general-purpose foundation models (FMs) or building "micro" domain-specific models [211]. An open challenge is to create a workflow to achieve a higher-level task, such as in industrial CPS. For example, a pipeline of activities in autonomous warehouse robotics could involve perception and sensing, prediction and reasoning, high-level planning, and human-in-the-loop oversight. The challenge is to shape this pipeline of multiple task-specific adapters into a unified, lightweight CPS application that can run on the edge.

- **Mid-term Goals (2–3 years):** One goal is to develop *parameter-efficient "functional patches"* (e.g., advanced LoRA variants) that allow a base CPS model to rapidly switch between distinct operational modes (e.g., from nominal driving to emergency hazard mitigation) without catastrophic forgetting. A promising direction is zero-shot adaptation to heterogeneous sensor resolutions and sampling rates, bridging the gap between pre-training and deployment. Rapidly adapting to environmental changes may also require data collection, so a promising direction is to explore planning and control models that can safely collect data in the presence of environmental changes.
- **Long-term Goals (≥ 5 years):** A worthy if challenging goal is to achieve *fully autonomous model evolution*, where the CPS identifies out-of-distribution (OOD) environments and self-synthesizes new adapters using local data and synthetic simulations. The ultimate goal is a *modular "plug-and-play" architecture* in which distributed CPS components can exchange and compose learned adapters to handle emergent, complex, multi-modal tasks in real-time.

2. Bidirectional Trust and Learning with Cognitive-Aware Human-CPS Teaming

Typically, trust in a resilient CPS is assumed to be a binary property: the human either trusts the system or does not. However, in the future, as resilient CPS are deployed more widely, there will be increasingly frequent settings in which the human user or operator is not an expert and will experience low-to-high cognitive failures while interacting with the system. Therefore, such systems should be able to accurately estimate human readiness and calibrate trust levels, thereby making trust a bi-directional property. This angle is intended to complement, rather than replace, ongoing research efforts to enable humans to assess the trustworthiness of the system under various scenarios. The same argument applies to learning: CPS should be designed so that the human operator can improve decision-making by interacting with the system, and vice versa, with the system improving by capturing tacit human (expert) knowledge and adapting over time. This is a critically required attribute for resilient CPS and will continue to grow in importance. It is only through such a calibrated and verified trust relationship that the human (or other interacting systems) can make decisions needed for the resilient operation of the CPS.

- **Mid-term Goals (2–3 years):** A goal is to design *real-time trust calibration frameworks* that use Explainable AI (XAI) to provide result-oriented explanations for tasks at various levels of autonomy. These should move beyond simple post-hoc reasoning to real-time explanations in the event of failures, allowing human operators to authorize effective, and occasionally high-risk, recovery strategies with full situational awareness and querying counterfactuals. The CPS must, over time, adapt through interactions with human experts, e.g., by adjusting its objective to capture implicit parameters not captured in the design.
- **Long-term Goals (≥ 5 years):** A long-term goal is to develop *cognitive-state-aware control regimes* where the CPS can autonomously delegate or reclaim authority based on inferred operator stress, fatigue, or low trust value for whatever reason. This includes resolving conflicts in multi-agent environments where teams of humans and

AI agents must maintain a shared mental model, including during critical operational periods such as cascading security attacks.

3. Provably Safe "Good Enough" Recovery Strategies

Resilience necessitates the ability to handle perturbations beyond pre-specified fault models, often requiring the system to settle for just "good enough" safety-critical functionality. This is a nod to the reality of complex CPS, where theoretically optimal recovery actions are not possible due to real-world constraints, such as policy or competitive constraints that disallow cooperation between some stakeholders, transient degraded connectivity between certain elements of the CPS, or the lack of capable human operators. Hence, we need to systematically develop the foundations for just-good-enough recovery. This should encompass several properties such as a notion of outcomes that absolutely must be avoided and then a graded degree of undesirability of other outcomes.

- **Mid-term Goals (2–3 years):** We have to refine *online safety verification components* (like Gatekeeper [7]) to handle highly nonlinear systems in rapidly changing environments. Research should focus on trajectories that guarantee a path to a safe state even when the planner explores unknown or unsafe state spaces.
- **Long-term Goals ($\geq$ 5 years):** We have to develop *deep recovery controllers* that can autonomously re-synthesize provable control logic in real-time in response to altered system dynamics. Such alterations may be caused by cyber or physical, failures or attacks. These strategies must be tamper-proof, utilizing hardware roots of trust (such as the ARM TrustZone), to ensure that the recovery process itself cannot be hijacked.

4. Verification of Neuro-Symbolic and Generative World Models

With the revolution of generative AI, it is inevitable that it will play a central role in CPS applications as well. To ensure that this can be used in critical CPS applications as well, it is important to ground generative AI in physical reality to prevent hallucinations in safety-critical settings. There is growing interest and promising results in neuro-symbolic models for CPS [125, 149]; neuro-symbolic models combine the learning power of neural networks with the structured reasoning of symbolic logic. Generative world models are used to predict how a physical environment (the "world") will evolve based on past observations and actions. There is a synergy between the two: neuro-symbolic models can be used to generate world representations due to their combined strengths of neural networks and symbolic logic. In that case, it would be important that the components of such models operate within provable safety and reliability bounds.

- **Mid-term Goals (2–3 years):** We have to advance neuro-symbolic specification languages (like NeSaL [237]) that allow verification engines to reason about high-dimensional pixel spaces, such as identifying a pedestrian in a blurry camera feed. This includes establishing probabilistic abstractions of perception that provide formal confidence intervals for learning-enabled CPS.
- **Long-term Goals ($\geq$ 5 years):** We have to create *physically interpretable world models* [172] that natively incorporate the underlying physics of the CPS components as well as the cyber request-response characteristics. We have to then enable closed-loop verification of such world models, where the system can prove its own resilience against real-world distribution shifts or other perturbations. A possible approach is coupling statistical models of perception with deterministic reachability analysis.

5. Resilient Interconnection for Large-Scale Infrastructure CPS

For infrastructure-scale CPS, resilience moves beyond individual component reliability to the system's capacity to serve society at a large scale. This requires managing widely varying connectivity among the different system elements. On a human level, this requires managing partially aligned incentives among different stakeholders and resulting in partial connectivity and cooperation among these stakeholders.

- **Mid-term Goals (2–3 years):** We have to develop *graph-theoretic measures* (e.g., algebraic connectivity, r-robustness, strong r-robustness) to design motion and communication controllers that preserve network integrity during coordinated movement or node failures. Research should focus on balancing planned network redundancy with cost and performance constraints, both during normal operation and under perturbations.
- **Long-term Goals ($\geq$ 5 years):** We have to develop methodologies to strengthen sub-systems within the CPS, taking into account the risks and the interdependencies among the assets. Models of the CPS system can help identify cost-effective resilience measures. However, these must not only model the technological elements, but also the economic (who are the stakeholders and what are their economic drivers) and the policy (what controls can each stakeholder implement and how can they collaborate) factors that guide CPS operational controls. These models, when instantiated with parameters from the real system, should enable rational and distributed decision-making among the multiple stakeholders on how to protect the CPS. Then, at runtime, based on inputs from sensors, the system can determine if a perturbation is currently underway and, if so, the optimal response to deploy. The goal is to ensure *system-wide resilience* (rather than individual subsystem or component-wise resilience) that can detect and adaptively react to both internal and exogenous perturbations.

In summary, when the five themes that we have described work together cohesively, they hold the greatest potential for us to achieve resilient CPS. Each theme has already delivered a rich array of solutions — and also presents to us a challenging set of questions that we, as an energized and well-coordinated community, will embark on and solve. We look forward to sustaining a vibrant field of resilient CPS research, working hand-in-hand with robust practical implementations and strong policy, all leading to widespread use of CPS in highly critical applications.

Acknowledgments

This material is based in part upon work supported by: (i) **S. Bagchi and H. Kim**: the DEVCOM ARL Army Research Office under Contract number W911NF-2020-221, and the National Science Foundation under Grant Numbers CNS-2333487 (CPS Frontier) and CNS-2038986; (ii) **J. Li and N. Li**: Cisco Research and the National Science Foundation under Grant Numbers CNS-2247794 and CNS-2207204; (iii) **S. Z. Yong**: National Science Foundation under Grant Numbers CNS-2312007 and CNS-2313814 and Office of Naval Research grant N00014-23-1-2093; (iv) **D. Panagou**: NSF Grant Numbers 1942907 (NSF CAREER), 2137195 (NSF I/UCRC), 2223845 (NSF CPS), and AFOSR Grant Number FA9550-23-1-0557 (Complex Networks); (v) **Y. Li**: DEVCOM ARL Army Research Office under Contract number W911NF-2020-221, NSF Grant Numbers IIS-2442739 (NSF CAREER), and CNS-2333491 (NSF-FRONTIER); (vi) **F. Kong**: the NSF Grant Numbers CNS-2442914 (NSF CAREER) and CNS-2333980 (NSF CPS); (vii) **H. Alemzadeh**: the NSF Grant Numbers CNS-2146295 (NSF CAREER) and CCF-2402941 (NSF SHF); (viii) **I. Ruchkin**: the NSF Grant Numbers CNS-2440920 (NSF CAREER) and CNS-2513076 (NSF CPS); (ix) **M. Ornik**: AFOSR Grant Number FA9550-23-1-0131, NASA Grant Number 80NSSC22M0070, and ONR Grant Numbers N00014-23-1-2651, N00014-23-1-2505, N00014-25-1-2369 and (x) **S. Mohan**: U.S. National Science Foundation grant CPS 2246937 (NSF CAREER). (xi) **A. Mukhopadhyay**: U.S. National Science Foundation grant CNS-2531369. (xii) **M.D. Lemmon and Y. Duan:** National Science Foundation CNS-2228092. (xiii) **M. Ma**: NSF Grant Numbers 2443803 (NSF CAREER) and 2427711 (NSF ReDDDoT). (xiv) **W. Luo**: NSF Grant Number 2528997 (NSF ERI). (xv) **Chaterji**: NSF Grant Numbers 2333487 (CPS Frontier) and 2146449 (CPS CAREER). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Fardin Abdi Taghi Abad, Renato Mancuso, Stanley Bak, Or Dantsker, and Marco Caccamo. 2016. Reset-based recovery for real-time cyber-physical systems with temporal safety constraints. In *2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 1–8.

[2] Fardin Abdi, Chien-Ying Chen, Monowar Hasan, Songran Liu, Sibin Mohan, and Marco Caccamo. 2018. Guaranteed Physical Security with Restart-Based Design for Cyber-Physical Systems. In *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS)*. 10–21. https://doi.org/10.1109/ICCPS.2018.00010

[3] Fardin Abdi, Chien-Ying Chen, Monowar Hasan, Songran Liu, Sibin Mohan, and Marco Caccamo. 2019. Preserving Physical Safety Under Cyber Attacks. *IEEE Internet of Things Journal* 6, 4 (2019), 6285–6300. https://doi.org/10.1109/JIOT.2018.2889866

[4] Aastha Acharya, Rebecca Russell, and Nisar R. Ahmed. 2022. Competency Assessment for Autonomous Agents using Deep Generative Models. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 8211–8218. https://doi.org/10.1109/IROS47612.2022.9981991 ISSN: 2153-0866.

[5] Guy Ackerson and K Fu. 1970. On state estimation in switching environments. *IEEE transactions on automatic control* 15, 1 (1970), 10–17.

[6] Ben Adcock and Nick Dexter. 2021. The gap between theory and practice in function approximation with deep neural networks. *SIAM Journal on Mathematics of Data Science* 3, 2 (2021), 624–655.

[7] Devansh Ramgopal Agrawal, Ruichang Chen, and Dimitra Panagou. 2024. gatekeeper: Online Safety Verification and Control for Nonlinear Systems in Dynamic Environments. *IEEE Transactions on Robotics* 40 (2024), 4358–4375. https://doi.org/10.1109/TRO.2024.3454415

[8] Devansh R Agrawal and Dimitra Panagou. 2021. Safe control synthesis via input constrained control barrier functions. In *2021 60th IEEE Conference on Decision and Control*. 6113–6118.

[9] Devansh R Agrawal, Hardik Parwana, Ryan K Cosner, Ugo Rosolia, Aaron D Ames, and Dimitra Panagou. 2021. A constructive method for designing safe multirate controllers for differentially-flat systems. *IEEE Control Systems Letters* 6 (2021), 2138–2143.

[10] Asrar Ahmed, Pradeep Varakantham, Yossiri Adulyasak, and Patrick Jaillet. 2013. Regret based robust solutions for uncertain Markov decision processes. In *27th International Conference on Neural Information Processing Systems*.

[11] R. Alur, P. Černý, and S. Zdancewic. 2006. Preserving Secrecy Under Refinement. In *Automata, Languages and Programming*. Springer Berlin Heidelberg, 107–118.

[12] R. Alur, T. Henzinger, G. Lafferriere, and G. J. Pappas. 2000. Discrete abstractions of hybrid systems. *Proc. IEEE* 88, 7 (July 2000), 971–984.

[13] R. Alur and T. A. Henzinger. 1990. Real-time Logics: Complexity and Expressiveness. In *LICS*. IEEE, 390–401.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[15] E. Asarin, O. Maler, and A. Pnueli. 1995. Reachability analysis of dynamical systems having piecewise-constant derivatives. *Theoretical Computer Science* 138 (1995), 35–66.

[16] Kıvanç Atilgan, Burak E Onuk, Pınar Köksal Coşkun, Fahri G Yeşil, Cemal Aslan, Abdullah Çolak, Aksüyek S Çelebi, and Hüseyin Bozbaş. 2021. Remote patient monitoring after cardiac surgery: the utility of a novel telemedicine system. *Journal of Cardiac Surgery* 36, 11 (2021), 4226–4234.

[17] Peter Auer, Thomas Jaksch, and Ronald Ortner. 2008. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems* 21 (2008).

[18] Alexis Aurandt, Phillip Jones, and Kristin Yvonne Rozier. 2022. Runtime Verification Triggers Real-time, Autonomous Fault Recovery on the CySat-I. In *Proceedings of the 14th NASA Formal Methods Symposium (NFM 2022) (Lecture Notes in Computer Science (LNCS), Vol. 13260)*. Springer, Cham, Caltech, California, USA.

[19] C. Baier and J. P. Katoen. 2008. *Principles of model checking*. The MIT Press.

[20] Ozan Baris, Yizhuo Chen, Gaofeng Dong, Liying Han, Tomoyoshi Kimura, Pengrui Quan, Ruijie Wang, Tianchen Wang, Tarek Abdelzaher, Mario Bergés, et al. 2025. Foundation Models for CPS-IoT: Opportunities and Challenges. *arXiv preprint arXiv:2501.16368* (2025).

[21] Ezio Bartocci. 2018. Monitoring, Learning and Control of Cyber-Physical Systems with STL (Tutorial). In *Runtime Verification*, Christian Colombo and Martin Leucker (Eds.). Springer International Publishing, Cham, 35–42.

[22] Calin Belta, Boyan Yordanov, and Ebru Aydin Gol. 2017. *Formal methods for discrete-time dynamical systems*. Vol. 89. Springer.

[23] Alessia Benevento, María Santos, Giuseppe Notarstefano, Kamran Paynabar, Matthieu Bloch, and Magnus Egerstedt. 2020. Multi-robot coordination for estimation and coverage of unknown spatial fields. In *2020 IEEE International Conference on Robotics and Automation*. 7740–7746.

[24] Melanie Berg and Kenneth A LaBel. 2016. Verification of triple modular redundancy (TMR) insertion for reliable and trusted systems. In *2016 MRQW Microelectronics Reliability and Qualification Working Meeting*.

[25] Mitchell Black, Ehsan Arabi, and Dimitra Panagou. 2021. A fixed-time stable adaptation law for safety-critical control under parametric uncertainty. In *2021 European Control Conference (ECC)*. IEEE, 1328–1333.

[26] Mitchell Black, Georgios Fainekos, Bardh Hoxha, Danil Prokhorov, and Dimitra Panagou. 2023. Safety under uncertainty: Tight bounds with risk-aware control barrier functions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 12686–12692.

[27] Kayla Boggess, Sarit Kraus, and Lu Feng. 2022. Toward policy explanations for multi-agent reinforcement learning. *arXiv preprint arXiv:2204.12568* (2022).

[28] Joseph Breeden and Dimitra Panagou. 2021. High relative degree control barrier functions under input constraints. In *2021 60th IEEE Conference on Decision and Control*. IEEE, 6119–6124.

[29] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[30] L Campo, P Mookerjee, and Y Bar-Shalom. 1991. State estimation for systems with sojourn-time-dependent Markov model switching. *IEEE Trans. Automat. Control* 36, 2 (1991), 238–243.

[31] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[32] C. Cassandras and S. Lafortune. 1999. *Introduction to discrete event systems*. Kluwer Academic Publishers, Boston, MA.

[33] Beatrice Cassottana, Muhammad M Roomi, Daisuke Mashima, and Giovanni Sansavini. 2023. Resilience analysis of cyber-physical systems: A review of models and methods. *Risk Analysis* 43, 11 (2023), 2359–2379.

[34] Matthew Cavorsi, Lorenzo Sabattini, and Stephanie Gil. 2023. Multirobot adversarial resilience using control barrier functions. *IEEE Transactions on Robotics* 40 (2023), 797–815.

[35] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317* (2017).

[36] Yash Chandak, Georgios Theocharous, Shiv Shankar, Sridhar Mahadevan, Martha White, and Philip S Thomas. 2020. Optimizing for the Future in Non-Stationary MDPs. *Thirty-seventh International Conference on Machine Learning (ICML)* (2020).

[37] Bingfeng Chen, Zhifeng Hao, Xiaofeng Cai, Ruichu Cai, Wen Wen, Jian Zhu, and Guangqiang Xie. 2019. Embedding Logic Rules Into Recurrent Neural Networks. *IEEE Access* 7 (2019). https://doi.org/10.1109/ACCESS.2019.2892140

[38] Chiao En Chen, Flavio Lorenzelli, Ralph E. Hudson, and Kung Yao. 2008. Stochastic maximum-likelihood DOA estimation in the presence of unknown nonuniform noise. *IEEE Transactions on Signal Processing* 56, 7 (2008), 3038–3044.

[39] Chien-Ying Chen, Debopam Sanyal, and Sibin Mohan. 2021. Indistinguishability Prevents Scheduler Side Channels in Real-Time Systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (Virtual Event, Republic of Korea) *(CCS '21)*. Association for Computing Machinery, New York, NY, USA, 666–684. https://doi.org/10.1145/3460120.3484769

[40] Hongkai Chen, Shan Lin, Scott A Smolka, and Nicola Paoletti. 2022. An STL-based formulation of resilience in cyber-physical systems. In *International Conference on Formal Modeling and Analysis of Timed Systems*. Springer, 117–135.

[41] Daniel M Cherenson, Devansh R Agrawal, and Dimitra Panagou. 2025. Autonomy Architectures for Safe Planning in Unknown Environments Under Budget Constraints. *arXiv preprint arXiv:2504.03001* (2025).

[42] Glen Chou, Dmitry Berenson, and Necmiye Ozay. 2020. Uncertainty-Aware Constraint Learning for Adaptive Safe Motion Planning from Demonstrations. In *4th Conference on Robot Learning, CoRL 2020, 16-18 November 2020, Virtual Event / Cambridge, MA, USA (Proceedings of Machine Learning Research, Vol. 155)*, Jens Kober, Fabio Ramos, and Claire J. Tomlin (Eds.). PMLR, 1612–1639. https://proceedings.mlr.press/v155/chou21a.html

[43] Glen Chou, Dmitry Berenson, and Necmiye Ozay. 2021. Learning constraints from demonstrations with grid and parametric representations. *Int. J. Robotics Res.* 40, 10-11 (2021). https://doi.org/10.1177/02783649211035177

[44] Glen Chou, Necmiye Ozay, and Dmitry Berenson. 2022. Safe Output Feedback Motion Planning from Images via Learned Perception Modules and Contraction Theory. In *Algorithmic Foundations of Robotics XV - Proceedings of the Fifteenth Workshop on the Algorithmic Foundations of Robotics, WAFR 2022, College Park, MD, USA, 22-24 June, 2022 (Springer Proceedings in Advanced Robotics, Vol. 25)*. Springer, 349–367.

[45] Michael R Clarkson and Fred B Schneider. 2010. Hyperproperties. *Journal of Computer Security* 18, 6 (2010), 1157–1210.

[46] Matthew Cleaveland, Pengyuan Lu, Oleg Sokolsky, Insup Lee, and Ivan Ruchkin. 2025. Conservative Perception Models for Probabilistic Verification. https://doi.org/10.48550/arXiv.2503.18077 arXiv:2503.18077 [cs] version: 2.

[47] Cyber-Physical Systems Public Working Group. 2017. *Framework for Cyber-Physical Systems: Volume 2, Working Group Reports*. NIST Special Publication 1500-202. National Institute of Standards and Technology (NIST), U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.1500-202 163 pp..

[48] James Bruster Dabney. 2021. Using Assume-Guarantee Contracts in Autonomous Spacecraft. Flight Software Workshop (FSW) Online: https://www.youtube.com/watch?v=zrtyiyNf674.

[49] James B. Dabney, Julia M. Badger, and Pavan Rajagopal. 2021. Adding a Verification View for an Autonomous Real-Time System Architecture. In *Proceedings of SciTech Forum (2021-0566)*. AIAA, Online. https://doi.org/10.2514/6.2021-0566

[50] James B Dabney, Julia M Badger, and Pavan Rajagopal. 2023. Trustworthy Autonomy for Gateway Vehicle System Manager. In *2023 IEEE Space Computing Conference (SCC)*. IEEE, 57–62.

[51] James Bruster Dabney, Pavan Rajagopal, and Julia M. Badger. 2022. Using Assume-Guarantee Contracts for Developmental Verification of Autonomous Spacecraft. Flight Software Workshop (FSW) Online: https://www.youtube.com/watch?v=HFnn6TzblPg.

[52] György Dán and Henrik Sandberg. 2010. Stealth attacks and protection schemes for state estimators in power systems. In *2010 first IEEE international conference on smart grid communications*. IEEE, 214–219.

[53] Giuseppe De Giacomo and Moshe Y. Vardi. 2013. Linear Temporal Logic and Linear Dynamic Logic on Finite Traces. In *23rd International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI Press, 854–860.

[54] Atsushi Deguchi et al. 2020. From smart city to society 5.0. *Society* 5 (2020), 43–65.

[55] Shirley Dion. 2013. *Global Precipitation Measurement (GPM) Safety Inhibit Timeline Tool.* Technical Report GSFC.ABS.7501.2012. NASA Goddard Space Flight Center, Greenbelt, MD, United States. https://ntrs.nasa.gov/citations/20130000831.

[56] Rahul Dixit, Ratne Babu Chinnam, and Harpreet Singh. 2020. Artificial intelligence and machine learning in sparse/inaccurate data situations. In *2020 IEEE Aerospace Conference*. IEEE.

[57] Alexandre Donzé. 2013. On Signal Temporal Logic. In *Runtime Verification*, Axel Legay and Saddek Bensalem (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 382–383.

[58] Viet Quoc Duong, Qiong Wu, Zhengyi Zhou, Eric Zavesky, WenLing Hsu, Han Zhao, and Huajie Shao. 2024. A General-Purpose Multi-Modal OOD Detection Framework. *Transactions on Machine Learning Research* (2024).

[59] Mica R Endsley. 2017. From here to autonomy: lessons learned from human–automation research. *Human factors* 59, 1 (2017), 5–27.

[60] H Erzberger and K Heere. 2010. Algorithm and operational concept for resolving short-range conflicts. *Proc. IMechE G J. Aerosp. Eng.* 224, 2 (2010), 225–243. https://doi.org/10.1243/09544100JAERO546 arXiv:http://pig.sagepub.com/content/224/2/225.full.pdf+html

[61] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R Endsley. 2008. The influence of agent reliability on trust in human-agent collaboration. In *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*. 1–8.

[62] Marco Faroni, Manuel Beschi, and Nicola Pedrocchi. 2022. Safety-aware time-optimal motion planning with uncertain human state estimation. *IEEE Robotics and Automation Letters* 7, 4 (2022), 12219–12226.

[63] Michael Fisher, Viviana Mascardi, Kristin Yvonne Rozier, Holger Schlingloff, Michael Winikoff, and Neil Yorke-Smith. 2021. Towards a Framework for Certification of Reliable Autonomous Systems. In *20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (Journal-first (JAAMAS) track)*. Springer.

[64] Sébastien Forestier, Rémy Portelas, Yoan Mollard, and Pierre-Yves Oudeyer. 2022. Intrinsically motivated goal exploration processes with automatic curriculum learning. *Journal of Machine Learning Research* 23, 152 (2022), 1–41.

[65] Dylan Foster and Max Simchowitz. 2020. Logarithmic regret for adversarial online control. In *International Conference on Machine Learning*. 3211–3221.

[66] Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable planning. *arXiv preprint arXiv:1709.10256* (2017).

[67] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*. 1587–1596.

[68] Kunal Garg, James Usevitch, Joseph Breeden, Mitchell Black, Devansh Agrawal, Hardik Parwana, and Dimitra Panagou. 2024. Advances in the theory of control barrier functions: Addressing practical challenges in safe control synthesis for autonomous and robotic systems. *Annual Reviews in Control* 57 (2024), 100945.

[69] Yuang Geng, Jake Brandon Baldauf, Souradeep Dutta, Chao Huang, and Ivan Ruchkin. 2024. Bridging Dimensions: Confident Reachability for High-Dimensional Controllers. In *Proc. of the International Symposium on Formal Methods (FM)*. Milano, Italy. https://doi.org/10.48550/arXiv.2311.04843 arXiv:2311.04843 [cs].

[70] A. Girard and G. J. Pappas. 2007. Approximation metrics for discrete and continuous systems. *IEEE Trans. Automat. Control* 25, 5 (May 2007), 782–798.

[71] A. Girard, G. Pola, and P. Tabuada. 2010. Approximately bisimilar symbolic models for incrementally stable switched systems. *IEEE Trans. Automat. Control* 55, 1 (January 2010), 116–126.

[72] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

[73] Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. 2016. Jointly embedding knowledge graphs and logical rules. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 192–202.

[74] David Ha and Jürgen Schmidhuber. 2018. World Models. In *Proc. of NeurIPS*. https://doi.org/10.5281/zenodo.1207631 arXiv:1803.10122 [cs, stat].

[75] Abigail Hammer, Matthew Cauwels, Benjamin Hertz, Phillip Jones, and Kristin Yvonne Rozier. 2021. Integrating Runtime Verification into an Automated UAS Traffic Management System. *Innovations in Systems and Software Engineering: A NASA Journal* (July 2021). https://doi.org/10.1007/s11334-021-00407-5

[76] Yuze He, Li Ma, Zhehao Jiang, Yi Tang, and Guoliang Xing. 2021. VI-eye: semantic-based 3D point cloud registration for infrastructure-assisted autonomous driving. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 573–586.

[77] Benjamin Hertz, Zachary Luppen, and Kristin Yvonne Rozier. 2021. Integrating Runtime Verification into a Sounding Rocket Control System. In *NASA Formal Methods - 13th International Symposium, NFM 2021, Virtual Event, May 24-28, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 12673)*, Aaron Dutle, Mariano M. Moscato, Laura Titolo, César A. Muñoz, and Ivan Perez (Eds.). Springer, 151–159. https://doi.org/10.1007/978-3-030-76384-8_10

[78] Erik Hollnagel, David D Woods, and Nancy Leveson. 2006. *Resilience engineering: Concepts and precepts*. Ashgate Publishing, Ltd.

[79] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2021. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[80] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[81] Susmit Jha, Ashish Tiwari, Sanjit Seshia, Tuhin Sahai, and Natarajan Shankar. 2017. TeLEx: Passive STL Learning Using Only Positive Examples. 208–224.

[82] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European conference on computer vision*. Springer, 709–727.

[83] Xu Jin, Wassim M Haddad, and Tansel Yucelen. 2017. An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems. *IEEE Trans. Automat. Control* 62, 11 (2017), 6058–6064.

[84] Chris Johannsen, Phillip Jones, Brian Kempa, Kristin Yvonne Rozier, and Pei Zhang. 2023. R2U2 Version 3.0: Re-Imagining a Toolchain for Specification, Resource Estimation, and Optimized Observer Generation for Runtime Verification in Hardware and Software. In *Computer Aided Verification*, Constantin Enea and Akash Lal (Eds.). Springer Nature Switzerland, Cham, 483–497.

[85] Rust John. 1988. Maximum likelihood estimation of discrete control processes. *SIAM Journal on Control and Optimization* 26, 5 (1988), 1006–1024.

[86] A. Jones, Z. Kong, and C. Belta. 2014. Anomaly detection in cyber-physical systems: A formal methods approach. In *53rd IEEE Conference on Decision and Control*. 848–853.

[87] Denizhan Kara, Tomoyoshi Kimura, Yatong Chen, Jinyang Li, Ruijie Wang, Yizhuo Chen, Tianshi Wang, Shengzhong Liu, and Tarek Abdelzaher. 2024. PhyMask: An Adaptive Masking Paradigm for Efficient Self-Supervised Learning in IoT. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. 97–111.

[88] Denizhan Kara, Tomoyoshi Kimura, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, and Tarek Abdelzaher. 2024. FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing. In *Proceedings of the ACM on Web Conference 2024*. 2795–2806.

[89] Ashish Kashinath, Disha Agarwala, Gabriel Kulp, Sourav Das, Sibin Mohan, and Radha Venkatagiri. 2025. Groundhog: A Restart-Based Systems Framework for Increasing Availability in Threshold Cryptosystems . In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 165–183. https://doi.org/10.1109/SP61157.2025.00056

[90] Sydney M. Katz, Anthony L. Corso, Christopher A. Strong, and Mykel J. Kochenderfer. 2022. Verification of Image-Based Neural Network Controllers Using Generative Models. *Journal of Aerospace Information Systems* 19, 9 (2022), 574–584. https://doi.org/10.2514/1.I011071 Publisher: American Institute of Aeronautics and Astronautics _eprint: https://doi.org/10.2514/1.I011071.

[91] Ramneet Kaur, Kaustubh Sridhar, Sangdon Park, Yahan Yang, Susmit Jha, Anirban Roy, Oleg Sokolsky, and Insup Lee. 2023. CODiT: Conformal out-of-distribution Detection in time-series data for cyber-physical systems. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023)*. 120–131.

[92] Brian Kempa, Pei Zhang, Phillip H. Jones, Joseph Zambreno, and Kristin Yvonne Rozier. 2020. Embedding Online Runtime Verification for Fault Disambiguation on Robonaut2. In *Proceedings of the 18th International Conference on Formal Modeling and Analysis of Timed Systems (FORMATS) (Lecture Notes in Computer Science (LNCS))*. Springer, Vienna, Austria, 196–214. http://research.temporallogic.org/papers/KZJZR20.pdf

[93] Nathaniel S Keplinger, Baiting Luo, Iliyas Bektas, Yunuo Zhang, Kyle Hollins Wray, Aron Laszka, Abhishek Dubey, and Ayan Mukhopadhyay. 2025. NS-Gym: Open-Source Simulation Environments and Benchmarks for Non-Stationary Markov Decision Processes. *arXiv preprint arXiv:2501.09646* (2025).

[94] Mohammad Khajenejad, Zeyuan Jin, Thach Ngoc Dinh, and Sze Zheng Yong. 2023. Resilient state estimation for nonlinear discrete-time systems via input and state interval observer synthesis. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 1826–1832.

[95] Mohammad Khajenejad and Sze Zheng Yong. 2022. Resilient state estimation and attack mitigation in cyber-physical systems. In *Security and Resilience in Cyber-Physical Systems: Detection, Estimation and Control*. Springer, 149–185.

[96] Sangjun Kim, Kyung-Joon Park, and Chenyang Lu. 2022. A survey on network security for cyber–physical systems: From threats to resilient design. *IEEE Communications Surveys & Tutorials* 24, 3 (2022), 1534–1573.

[97] Taekyung Kim, Robin Inho Kee, and Dimitra Panagou. 2025. Learning to Refine Input Constrained Control Barrier Functions via Uncertainty-Aware Online Parameter Adaptation. In *2025 IEEE International Conference on Robotics and Automation*.

[98] Taekyung Kim and Dimitra Panagou. 2024. Visibility-Aware RRT* for Safety-Critical Navigation of Perception-Limited Robots in Unknown Environments. *arXiv preprint arXiv:2406.07728, accepted in RA-L, in press; will be presented in IROS 2025* (2024).

[99] Tomoyoshi Kimura, Jinyang Li, Tianshi Wang, Yizhuo Chen, Ruijie Wang, Denizhan Kara, Maggie Wigness, Joydeep Bhattacharyya, Mudhakar Srivatsa, Shengzhong Liu, et al. 2024. Vibrofm: Towards micro foundation models for robust multimodal iot sensing. In *2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*. IEEE, 10–18.

[100] Tomoyoshi Kimura, Ashitabh Misra, Yizhuo Chen, Denizhan Kara, Jinyang Li, Tianshi Wang, Ruijie Wang, Joydeep Bhattacharyya, Jae Kim, Prashant Shenoy, et al. 2024. The Case for Micro Foundation Models to Support Robust Edge Intelligence. In *2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 23–31.

[101] M. Kloetzer and C. Belta. 2008. A fully automated framework for control of linear systems from temporal logic specifications. *IEEE Trans. Automat. Control* 53, 1 (2008), 287–297.

[102] Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *European conference on machine learning*. Springer, 282–293.

[103] R. Kumar and V.K. Garg. 1995. *Modeling and Control of Logical Discrete Event Systems*. Kluwer Academic Publishers.

[104] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1931–1941.

[105] Y. Lai et al. 2025. Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & Society* (2025).

[106] Pat Langley. 2016. Explainable agency in human-robot interaction. In *AAAI fall symposium series*.

[107] Heath J LeBlanc, Haotian Zhang, Xenofon Koutsoukos, and Shreyas Sundaram. 2013. Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications* 31, 4 (2013), 766–781.

[108] Erwan Lecarpentier and Emmanuel Rachelson. 2019. Non-stationary Markov decision processes, a worst-case approach using model-based reinforcement learning. *Advances in neural information processing systems* 32 (2019).

[109] Haejoon Lee and Dimitra Panagou. 2025. Maintaining Strong r-Robustness in Reconfigurable Multi-Robot Networks using Control Barrier Functions. In *2025 IEEE International Conference on Robotics and Automation*.

[110] Insup Lee and Oleg Sokolsky. 2010. Medical cyber physical systems. In *Proceedings of the 47th Design Automation Conference* (Anaheim, California) *(DAC '10)*. Association for Computing Machinery, New York, NY, USA, 743–748. https://doi.org/10.1145/1837274.1837463

[111] Insup Lee, Oleg Sokolsky, Sanjian Chen, John Hatcliff, Eunkyoung Jee, BaekGyu Kim, Andrew King, Margaret Mullen-Fortino, Soojin Park, Alexander Roederer, et al. 2011. Challenges and research directions in medical cyber–physical systems. *Proc. IEEE* 100, 1 (2011), 75–90.

[112] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

[113] Nancy Leveson. 2004. A new accident model for engineering safer systems. *Safety science* 42, 4 (2004), 237–270.

[114] Nancy G. Leveson. 2011. *Engineering a safer world: Systems thinking applied to safety*. MIT Press, Cambridge, MA.

[115] Jianwen Li, Moshe Y. Vardi, and Kristin Y. Rozier. 2019. Satisfiability Checking for Mission-Time LTL. In *Proceedings of 31st International Conference on Computer Aided Verification (CAV) (LNCS, Vol. 11562)*. Springer, New York, NY, USA, 3–22.

[116] X. Li et al. 2025. Trust-Triggered Cyber-Physical-Human System for Human-Robot Collaboration in Flexible Manufacturing. *IEEE Transactions on Industrial Informatics* (2025). Early Access.

[117] Yitao Liang and Guy Van den Broeck. 2019. Learning logistic circuits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4277–4286.

[118] Dongxin Liu, Tianshi Wang, Shengzhong Liu, Ruijie Wang, Shuochao Yao, and Tarek Abdelzaher. 2021. Contrastive self-supervised representation learning for sensing signals from the time-frequency perspective. In *2021 International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–10.

[119] Mengyu Liu, Lin Zhang, Vir V Phoha, and Fanxin Kong. 2023. Learn-to-respond: Sequence-predictive recovery from sensor attacks in cyber-physical systems. In *2023 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 78–91.

[120] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. 2024. FOCAL: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. *Advances in Neural Information Processing Systems* 36 (2024).

[121] S. Liu, A. Trivedi, X. Yin, and **M. Zamani**. 2022. Secure-by-construction synthesis of cyber-physical systems. *Annual Reviews in Control* (2022).

[122] Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. 2025. PAVE: Patching and Adapting Video Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3306–3317.

[123] Chen-Yi Lu, Kasra Derakhshandeh, and Somali Chaterji. 2025. Improving semi-supervised semantic segmentation with sliced-Wasserstein feature alignment and uniformity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 20233–20243.

[124] Pengyuan Lu, Lin Zhang, Mengyu Liu, Kaustubh Sridhar, Oleg Sokolsky, Fanxin Kong, and Insup Lee. 2024. Recovery from adversarial attacks in cyber-physical systems: Shallow, deep, and exploratory works. *Comput. Surveys* 56, 8 (2024), 1–31.

[125] Zhen Lu, Imran Afridi, Hong Jin Kang, Ivan Ruchkin, and Xi Zheng. 2024. Surveying neuro-symbolic approaches for reliable artificial intelligence of things. *Journal of Reliable Intelligent Environments* 10, 3 (2024), 257–279.

[126] J. Lunze, B. Nixdorf, and J. Schröder. 1999. Deterministic discrete-event representation of continuous-variable systems. *Automatica* 35 (March 1999), 395–406.

[127] Baiting Luo, Shreyas Ramakrishna, Ava Pettet, Christopher Kuhn, Gabor Karsai, and Ayan Mukhopadhyay. 2023. Dynamic Simplex: Balancing Safety and Performance in Autonomous Cyber Physical Systems. In *Proceedings of the ACM/IEEE 14th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2023) (ICCPS '23)*. Association for Computing Machinery, New York, NY, USA, 177–186. https://doi.org/10.1145/3576841.3585934

[128] Baiting Luo, Yunuo Zhang, Abhishek Dubey, and Ayan Mukhopadhyay. 2024. Act as you learn: Adaptive decision-making in non-stationary markov decision processes. *International Conference on Autonomous Agents and Multi-Agent Systems* (2024).

[129] Wenhao Luo, Nilanjan Chakraborty, and Katia Sycara. 2020. Minimally disruptive connectivity enhancement for resilient multi-robot teams. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11809–11816.

[130] Wenhao Luo and Katia Sycara. 2019. Minimum k-connectivity maintenance for robust multi-robot systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7370–7377.

[131] Wenhao Luo, Sha Yi, and Katia Sycara. 2020. Behavior mixing with minimum global and subgroup connectivity maintenance for large-scale multi-robot systems. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9845–9851.

[132] Zachary A. Luppen, Dae Young Lee, and Kristin Yvonne Rozier. 2021. A Case Study in Formal Specification and Runtime Verification of a CubeSat Communications System. In *SciTech*. AIAA, Nashville, TN, USA. https://doi.org/10.2514/6.2021-0997.c1

[133] Meiyi Ma, Sarah Preum, Mohsin Ahmed, William Tärneberg, Abdeltawab Hendawi, and John Stankovic. 2019. Data sets, modeling, and decision making in smart cities: A survey. *ACM Transactions on Cyber-Physical Systems* 4, 2 (2019), 1–28. https://doi.org/10.1145/3355283

[134] Meiyi Ma, John Stankovic, Ezio Bartocci, and Lu Feng. 2021. Predictive monitoring with logic-calibrated uncertainty for cyber-physical systems. *ACM Transactions on Embedded Computing Systems* 20, 5s (2021), 1–25.

[135] P. Madhusudan, W. Nam, and R. Alur. 2003. Symbolic Computational Techniques for Solving Games. *Electronic Notes in Theoretical Computer Science* 89, 4 (2003).

[136] Kanak Mahadik, Christopher Wright, Jinyi Zhang, Milind Kulkarni, Saurabh Bagchi, and Somali Chaterji. 2016. Sarvavid: a domain specific language for developing scalable computational genomics applications. In *Proceedings of the 2016 International Conference on Supercomputing (ICS)*. 1–12.

[137] O. Maler and D. Nickovic. 2004. Monitoring temporal properties of continuous signals. In *Proceedings of Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems (FORMATS)*. 152–166.

[138] O. Maler, A. Pnueli, and J. Sifakis. 1995. On the synthesis of discrete controllers for timed systems. In *Symposium on Theoretical Aspects of Computer Science (LNCS, Vol. 900)*, E. W. Mayr and C. Puech (Eds.). Springer-Verlag, 229–242.

[139] Zhenjiang Mao, Carson Sobolewski, and Ivan Ruchkin. 2024. How Safe Am I Given What I See? Calibrated Prediction of Safety Chances for Image-Controlled Autonomy. In *Proc. of the Annual Conference on Learning for Dynamics and Control (L4DC)*. https://doi.org/10.48550/arXiv.2308.12252 arXiv:2308.12252 [cs].

[140] Rahul Umesh et al. Mhapsekar. 2024. Building Trust in AI-Driven Decision Making for Cyber-Physical Systems (CPS): A Comprehensive Review. *arXiv preprint arXiv:2405.06347* (2024).

[141] R. Milner. 1989. *Communication and Concurrency*. Prentice-Hall, Inc.

[142] Shaunak Mishra, Yasser Shoukry, Nikhil Karamchandani, Suhas N Diggavi, and Paulo Tabuada. 2016. Secure state estimation against sensor attacks in the presence of noise. *IEEE Transactions on Control of Network Systems* 4, 1 (2016), 49–59.

[143] Sayan Mitra, Corina Păsăreanu, Pavithra Prabhakar, Sanjit A. Seshia, Ravi Mangal, Yangge Li, Christopher Watson, Divya Gopinath, and Huafeng Yu. 2025. Formal Verification Techniques for Vision-Based Autonomous Systems – A Survey. In *Principles of Verification: Cycling the Probabilistic Landscape : Essays Dedicated to Joost-Pieter Katoen on the Occasion of His 60th Birthday, Part III*, Nils Jansen, Sebastian Junges, Benjamin Lucien Kaminski, Christoph Matheja, Thomas Noll, Tim Quatmann, Mariëlle Stoelinga, and Matthias Volk (Eds.). Springer Nature Switzerland, Cham, 89–108. https://doi.org/10.1007/978-3-031-75778-5_5

[144] Yilin Mo and Bruno Sinopoli. 2010. False data injection attacks in control systems. In *Preprints of the 1st workshop on Secure Control Systems*, Vol. 1.

[145] Hossein Mohammadi Rouzbahani, Hadis Karimipour, Abolfazl Rahimnejad, Ali Dehghantanha, and Gautam Srivastava. 2020. Anomaly detection in cyber-physical systems using machine learning. *Handbook of big data privacy* (2020), 219–235.

[146] Luis Montestruque and MD Lemmon. 2015. Globally coordinated distributed storm water management system. In *Proceedings of the 1st ACM International Workshop on Cyber-Physical Systems for Smart Water Networks*. 1–6.

[147] T. Moor and J. Raisch. 1999. Supervisory control of hybrid systems within a behavioural framework. *Systems & Control Letters* 38, 3 (1999), 157–166.

[148] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15991–16111.

[149] Md Shirajum Munir, Ki Tae Kim, Apurba Adhikary, Walid Saad, Sachin Shetty, Seong-Bae Park, and Choong Seon Hong. 2023. Neuro-symbolic explainable artificial intelligence twin for zero-touch IoE in wireless network. *IEEE Internet of Things Journal* 10, 24 (2023), 22451–22468.

[150] Vishnu Murali, Ashutosh Trivedi, and Majid Zamani. 2023. Co-Buchi Barrier Certificates for Discrete-time Dynamical Systems. *arXiv preprint arXiv:2311.07695* (2023).

[151] Carlos Murguia and Justin Ruths. 2016. Cusum and chi-squared attack detection of compromised sensors. In *2016 IEEE Conference on Control Applications (CCA)*. IEEE, 474–480.

[152] Dietmar P.F. Möller and Hamid Vakilzadian. 2016. Cyber-physical systems in smart transportation. In *2016 IEEE International Conference on Electro Information Technology (EIT)*. 0776–0781. https://doi.org/10.1109/EIT.2016.7535338

[153] Kaleb Ben Naveed, An Dang, Rahul Kumar, and Dimitra Panagou. 2024. meSch: Multi-Agent Energy-Aware Scheduling for Task Persistence. *arXiv preprint arXiv:2406.04560, to be presented in IROS 2025* (2024).

[154] Farhad Nawaz and Melkior Ornik. 2020. Explorative probabilistic planning with unknown target locations. In *59th IEEE Conference on Decision and Control*. 2732–2737.

[155] NVIDIA. 2018. NVIDIA Drive Simulation. https://www.nvidia.com/en-us/self-driving-cars/drive-constellation/. (2018).

[156] Naoko Okubo and Tsutomu Kobayashi. 2020. Using R2U2 in JAXA program. Electronic correspondence; https://www.esa.int/Enabling_Support/Operations/OPS-SAT. Series of emails and zoom call from JAXA to PI with technical questions about embedding R2U2 into an autonomous satellite mission with a provable memory bound of 200KB.

[157] Pio Ong, Beatrice Capelli, Lorenzo Sabattini, and Jorge Cortés. 2023. Nonsmooth control barrier function design of continuous constraints for network connectivity maintenance. *Automatica* 156 (2023), 111209.

[158] Melkior Ornik, Steven Carr, Arie Israel, and Ufuk Topcu. 2019. Myopic control of systems with unknown dynamics. In *American Control Conference*. 1064–1071.

[159] Melkior Ornik, Jie Fu, Niklas T. Lauffer, W. K. Perera, Mohammed Alshiekh, Masahiro Ono, and Ufuk Topcu. 2018. Expedited learning in MDPs with side information. In *57th IEEE Conference on Decision and Control*. 1941–1948.

[160] Melkior Ornik and Ufuk Topcu. 2021. Learning and planning for time-varying MDPs using maximum likelihood estimation. *Journal of Machine Learning Research* 22 (2021).

[161] J. Ouaknine and J. Worrell. 2008. Some Recent Results in Metric Temporal Logic. In *Formal Modeling and Analysis of Timed Systems*, Franck Cassez and Claude Jard (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–13.

[162] Ram Padmanabhan and Melkior Ornik. 2025. Energetic resilience of linear driftless systems. In *11th IFAC Symposium on Robust Control Design*.

[163] Manuel Ramón Castillo Padrós, Nuria Pastor, Júlia Altarriba Paracolls, Marcelino Mosquera Peña, Denise Pergolizzi, and Àngels Salvador Vergès. 2023. A smart system for remote monitoring of patients in palliative care (humanITcare platform): mixed methods study. *JMIR formative research* 7, 1 (2023), e45654.

[164] Miroslav Pajic, Insup Lee, and George J Pappas. 2016. Attack-resilient state estimation for noisy dynamical systems. *IEEE Transactions on Control of Network Systems* 4, 1 (2016), 82–92.

[165] Jiayi Pan, Glen Chou, and Dmitry Berenson. 2023. Data-Efficient Learning of Natural Language to Linear Temporal Logic Translators for Robot Task Specification. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*. IEEE, 11554–11561. https://doi.org/10.1109/ICRA48891.2023.10161125

[166] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (2000), 286–297.

[167] Guy Paré, Mirou Jaana, and Claude Sicotte. 2007. Systematic review of home telemonitoring for chronic diseases: the evidence base. *Journal of the American Medical Informatics Association* 14, 3 (2007), 269–277.

[168] D. M. R. Park. London, 1981. Concurrency and automata on infinite sequences. *in Proceedings of the 5th GI Conference on Theoretical Computer Science, LNCS* 104 (London, 1981), 167–183.

[169] Pablo A. Parrilo. 2003. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming* 96 (2003), 293–320.

[170] Corina S. Pasareanu, Ravi Mangal, Divya Gopinath, Sinem Getir Yaman, Calum Imrie, Radu Calinescu, and Huafeng Yu. 2023. Closed-Loop Analysis of Vision-Based Autonomous Systems: A Case Study. In *Computer Aided Verification (Lecture Notes in Computer Science)*. Springer Nature Switzerland, Cham, 289–303. https://doi.org/10.1007/978-3-031-37706-8_15

[171] Jason Pazis and Ronald Parr. 2016. Efficient PAC-optimal exploration in concurrent, continuous state MDPs with delayed updates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[172] Jordan Peper, Zhenjiang Mao, Yuang Geng, Siyuan Pan, and Ivan Ruchkin. 2025. Four Principles for Physically Interpretable World Models. In *Proc. of 2nd International Conference on Neuro-symbolic Systems (NeuS)*. PMLR, Philadelphia, PA, USA. https://doi.org/10.48550/arXiv.2503.02143

[173] Jordan Peper, Yan Miao, Sayan Mitra, and Ivan Ruchkin. 2025. Towards Unified Probabilistic Verification and Validation of Vision-Based Autonomy. In *Automated Technology for Verification and Analysis*. Springer Nature Switzerland, Cham, 231–259. https://doi.org/10.1007/978-3-032-08707-2_11

[174] Ava Pettet, Yunuo Zhang, Baiting Luo, Kyle Wray, Hendrik Baier, Aron Laszka, Abhishek Dubey, and Ayan Mukhopadhyay. 2024. Decision Making in Non-Stationary Environments with Policy-Augmented Search. *arXiv preprint arXiv:2401.03197* (2024).

[175] Sandro Pinto and Nuno Santos. 2019. Demystifying arm trustzone: A comprehensive survey. *ACM computing surveys (CSUR)* 51, 6 (2019), 1–36.

[176] Mohammad Pirani, Aritra Mitra, and Shreyas Sundaram. 2023. Graph-theoretic approaches for analyzing the resilience of distributed control systems: A tutorial and survey. *Automatica* 157 (2023), 111264. https://doi.org/10.1016/j.automatica.2023.111264

[177] G. Pola, A. Girard, and P. Tabuada. 2008. Approximately bisimilar symbolic models for nonlinear control systems. *Automatica* 44, 10 (October 2008), 2508–2516.

[178] Stephen Prajna and Ali Jadbabaie. 2004. Safety verification of hybrid systems using barrier certificates. In *Hybrid Systems: Computation and Control (Lecture Notes in Computer Science)*. 477–492. https://doi.org/10.1007/978-3-540-24743-2_32

[179] Amanda Prorok, Matthew Malencia, Luca Carlone, Gaurav S Sukhatme, Brian M Sadler, and Vijay Kumar. 2021. Beyond robustness: A taxonomy of approaches towards resilient multi-robot systems. *arXiv preprint arXiv:2109.12343* (2021).

[180] G. Puthumanaillam, X. Liu, N. Mehr, and M. Ornik. 2024. Weathering ongoing uncertainty: Learning and planning in a time-varying partially observable environment. In *2024 IEEE International Conference on Robotics and Automation*.

[181] Gokul Puthumanaillam, Yuvraj Mamik, and Melkior Ornik. 2024. Online Learning and Planning in Time-Varying Environments: An Aircraft Case Study. In *AIAA SCITECH 2024 Forum*.

[182] Gokul Puthumanaillam, Paolo Padrao, Jose Fuentes, Leonardo Bobadilla, and Melkior Ornik. 2025. Enhancing Robot Navigation Policies with Task-Specific Uncertainty Management. In *24th International Conference on Autonomous Agents and Multiagent Systems*.

[183] Gokul Puthumanaillam, Jae Hyuk Song, Nurzhan Yesmagambet, Shinkyu Park, and Melkior Ornik. 2025. TAB-Fields: A maximum entropy framework for mission-aware adversarial planning. In *7th Annual Learning for Dynamics & Control Conference*.

[184] Hang Qiu, Po-Han Huang, Namo Asavisanu, Xiaochen Liu, Konstantinos Psounis, and Ramesh Govindan. 2022. AutoCast: scalable infrastructure-less cooperative perception for distributed collaborative driving. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 128–141.

[185] Yuning Qiu, Teruhisa Misu, and Carlos Busso. 2022. Unsupervised Scalable Multimodal Driving Anomaly Detection. *IEEE Transactions on Intelligent Vehicles* (2022).

[186] Rahul Rai and Chandan K Sahu. 2020. Driven by data or derived through physics? a review of hybrid physics guided machine learning techniques with cyber-physical system (cps) focus. *IEEe Access* 8 (2020), 71050–71073.

[187] J. Raisch and S. D. O'Young. 1998. Discrete approximation and supervisory control of continuous systems. *IEEE Trans. Automat. Control* 43, 4 (April 1998), 569–573.

[188] Shreyas Ramakrishna, Zahra Rahiminasab, Gabor Karsai, Arvind Easwaran, and Abhishek Dubey. 2022. Efficient out-of-distribution detection using latent space of β-vae for cyber-physical systems. *ACM Transactions on Cyber-Physical Systems (TCPS)* 6, 2 (2022), 1–34.

[189] Md Masud Rana and Kamal Hossain. 2023. Connected and autonomous vehicles and infrastructures: A literature review. *International Journal of Pavement Research and Technology* 16, 2 (2023), 264–284.

[190] Denise Ratasich, Faiq Khalid, Florian Geissler, Radu Grosu, Muhammad Shafique, and Ezio Bartocci. 2019. A roadmap toward the resilient internet of things for cyber-physical systems. *IEEE Access* 7 (2019), 13260–13283.

[191] Thomas Reinbacher, Kristin Y. Rozier, and Johann Schumann. 2014. Temporal-Logic Based Runtime Observer Pairs for System Health Management of Real-Time Systems. In *Proceedings of the 20th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS) (Lecture Notes in Computer Science (LNCS), Vol. 8413)*. Springer-Verlag, 357–372.

[192] G. Reißig, A. Weber, and M. Rungger. 2017. Feedback refinement relations for the synthesis of symbolic controllers. *IEEE Trans. Automat. Control* 62, 4 (2017), 1781–1796.

[193] Kristin Yvonne Rozier. 2016. Specification: The Biggest Bottleneck in Formal Methods and Autonomy. In *Proceedings of 8th Working Conference on Verified Software: Theories, Tools, and Experiments (VSTTE 2016) (LNCS, Vol. 9971)*. Springer-Verlag, Toronto, ON, Canada, 1–19. https://doi.org/10.1007/978-3-319-48869-1_2

[194] Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. 2015. Policy Distillation. https://doi.org/10.48550/ARXIV.1511.06295

[195] Jason Ryan, Mary Cummings, Nick Roy, Ashis Banerjee, and Axel Schulte. 2011. Designing an interactive local and global decision support system for aircraft carrier deck scheduling. In *Infotech@Aerospace*. AIAA.

[196] Sahar Salehi, Alireza Olyaeemanesh, Mohammadreza Mobinizadeh, Ensieh Nasli-Esfahani, and Hossein Riazi. 2020. Assessment of remote patient monitoring (RPM) systems for patients with type 2 diabetes: a systematic review and meta-analysis. *Journal of Diabetes & Metabolic Disorders* 19 (2020), 115–127.

[197] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

[198] Mariana Segovia-Ferreira, Jose Rubio-Hernan, Ana Cavalli, and Joaquin Garcia-Alfaro. 2024. A survey on cyber-resilience approaches for cyber-physical systems. *Comput. Surveys* 56, 8 (2024), 1–37.

[199] Taha Shafa and Melkior Ornik. 2022. Reachability of Nonlinear Systems with Unknown Dynamics. *IEEE Trans. Automat. Control* (2022).

[200] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*. 492–504.

[201] Hao Shao, Yuxuan Wang, Ruo-Ping Chen, et al. 2024. LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[202] Yuan Shen, Bhargav Chandaka, Zhi-Hao Lin, Albert Zhai, Hang Cui, David Forsyth, and Shenlong Wang. 2023. Sim-on-wheels: physical world in the loop simulation for self-driving. *IEEE Robotics and Automation Letters* 8, 12 (2023), 8192–8199.

[203] Thomas B. Sheridan. 2014. *Humans and Automation: System Design and Research Issues*. Wiley.

[204] Shuyao Shi, Jiahe Cui, Zhehao Jiang, Zhenyu Yan, Guoliang Xing, Jianwei Niu, and Zhenchao Ouyang. 2022. VIPS: Real-time perception fusion for infrastructure-assisted autonomous driving. In *Proceedings of the 28th annual international conference on mobile computing and networking (MobiCom)*. 133–146.

[205] Rohan Sinha, Amine Elhafsi, Christopher Agia, Matthew Foutter, Edward Schmerling, and Marco Pavone. 2024. Real-time anomaly detection and reactive planning with large language models. *arXiv preprint arXiv:2407.08735* (2024).

[206] S Sreedharan, T Chakraborti, and S Kambhampati. 2020. The emerging landscape of explainable automated planning & decision making. IJCAI.

[207] J. Stray et al. 2025. Preliminary Quantitative Study on Explainability and Trust in AI Systems. *arXiv preprint arXiv:2510.15769* (2025).

[208] Lei Sun, Kailun Yang, Xinxin Hu, Weijian Hu, and Kaiwei Wang. 2020. Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. *IEEE Robotics and Automation Letters* 5, 4 (2020), 5558–5565.

[209] P. Tabuada. 2009. *Verification and Control of Hybrid Systems: A symbolic approach*. Springer.

[210] Paulo Tabuada. 2009. *Verification and control of hybrid systems: a symbolic approach*. Springer Science & Business Media.

[211] Jianhua Tang, Jiao Chen, Jiayi He, Fangfang Chen, Zuohong Lv, Guangjie Han, Zuozhu Liu, Howard H Yang, and Weihua Li. 2025. Towards general industrial intelligence: A survey of large models as a service in industrial IoT. *IEEE Communications Surveys & Tutorials* (2025).

[212] Nathan et al. Tenhundfeld. 2022. Assessment of Trust in Automation in the "Real World": Requirements for New Trust in Automation Measurement Techniques for Use by Practitioners. *Human Factors* 64, 4 (2022), 534–555.

[213] Pranay Thangeda and Melkior Ornik. 2020. PROTRIP: Probabilistic risk-aware optimal transit planner. In *23rd IEEE International Conference on Intelligent Transportation Systems*.

[214] Pranay Thangeda and Melkior Ornik. 2022. Adaptive sampling site selection for robotic exploration in unknown environments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4120–4125.

[215] Pranay Thangeda, Melkior Ornik, and Ufuk Topcu. 2022. Expedited Online Learning with Spatial Side Information. *IEEE Trans. Automat. Control* (2022).

[216] W. Thomas. 1995. On the synthesis of strategies in infinite games. In *Proceedings of the 12th Annual Symposium on Theoretical Aspects of Computer Science (LNCS, Vol. 900)*, E. W. Mayr and C. Puech (Eds.). Springer Berlin Heidelberg, 1–13.

[217] Shuo Tian, Wenbo Yang, Jehane Michael Le Grange, Peng Wang, Wei Huang, and Zhewei Ye. 2019. Smart healthcare: making medical care more intelligent. *Global Health Journal* 3, 3 (2019), 62–65.

[218] James Usevitch and Dimitra Panagou. 2021. Resilient trajectory propagation in multirobot networks. *IEEE Trans. on Robotics* 38, 1 (2021), 42–56.

[219] Niraj Varma, Frieder Braunschweig, Haran Burri, Gerhard Hindricks, Dominik Linz, Yoav Michowitz, Renato Pietro Ricci, and Jens Cosedis Nielsen. 2023. Remote monitoring of cardiac implantable electronic devices and disease management. *Europace* 25, 9 (2023), euad233.

[220] Thomas Waite, Yuang Geng, Trevor Turnquist, Ivan Ruchkin, and Radoslav Ivanov. 2025. State-Dependent Conformal Perception Bounds for Neuro-Symbolic Verification of Autonomous Systems. In *Proc. of 2nd International Conference on Neuro-symbolic Systems (NeuS)*. PMLR, Philadelphia, PA, USA. https://doi.org/10.48550/arXiv.2502.21308

[221] LeiChen Wang, Simon Giebenhain, Carsten Anklam, and Bastian Goldluecke. 2021. Radar ghost target detection via multimodal transformers. *IEEE Robotics and Automation Letters* 6, 4 (2021), 7758–7765.

[222] Tianshi Wang, Yizhuo Chen, Qikai Yang, Dachun Sun, Ruijie Wang, Jinyang Li, Tomoyoshi Kimura, and Tarek Abdelzaher. 2024. Data augmentation for human activity recognition via condition space interpolation within a generative model. In *2024 33rd International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–9.

[223] Tianshi Wang, Jinyang Li, Ruijie Wang, Denizhan Kara, Shengzhong Liu, Davis Wertheimer, Antoni Viros i Martin, Raghu Ganti, Mudhakar Srivatsa, and Tarek Abdelzaher. 2023. Sudokusens: Enhancing deep learning robustness for iot sensing applications using a generative approach. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 15–27.

[224] Tianshi Wang, Jinyang Li, Qikai Yang, Ruijie Wang, Yizhuo Chen, Dachun Sun, Bohan Li, Yigong Hu, Tomoyoshi Kimura, Denizhan Kara, and Tarek Abdelzaher. 2025. DynaGen: Conditional Diffusion Models for Enhancing Acoustic and Seismic-Based Vehicle Detection. In *In Proc. IEEE Conference on Computer Communications (Infocom)* (London, UK).

[225] Xinyi Wang, Taekyung Kim, Bardh Hoxha, Georgios Fainekos, and Dimitra Panagou. 2025. Safe Navigation in Uncertain Crowded Environments Using Risk Adaptive CVaR Barrier Functions. In *2025 International Conference on Intelligent Robots and Systems*.

[226] Y. Wang et al. 2024. Trust-Aware Reflective Control for Fault-Resilient Dynamic Task Response in Human–Swarm Cooperation. *Robotics* 5, 1 (2024), 22.

[227] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys* (2020).

[228] Z. Wang et al. 2024. Enhancing Human–Machine Collaboration: A Trust-Aware Trajectory Planning Framework for Assistive Aerial Teleoperation. *Aerospace* 13, 9 (2024), 876.

[229] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

[230] Yuxi Wei, Zi Wang, Yifan Lu, et al. 2024. ChatSim: Editable Scene Simulation for Autonomous Driving via Collaborative LLM-Agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[231] Jessica L. Wildman, T. Nguyen, et al. 2024. Trust in Human-Agent Teams: A Multilevel Perspective and Future Research Agenda. *Organizational Psychology Review* 14, 3 (2024).

[232] J. C. Willems. 2007. The behavioral approach to open and interconnected systems. *IEEE Control Systems Magazine* 27, 6 (2007), 46–99.

[233] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16.

[234] Eric M. Wolff, Ufuk Topcu, and Richard M. Murray. 2012. Robust control of uncertain Markov Decision Processes with temporal logic specifications. In *51st IEEE Conference on Decision and Control*. 3372–3379.

[235] Haoze Wu, Teruhiro Tagomori, Alexander Robey, Fengjun Yang, Nikolai Matni, George Pappas, Hamed Hassani, Corina Pasareanu, and Clark Barrett. 2023. Toward Certified Robustness Against Real-World Distribution Shifts. *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (Feb. 2023), 537–553. https://doi.org/10.1109/SaTML54575.2023.00042 Conference Name: 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) ISBN: 9781665462990 Place: Raleigh, NC, USA Publisher: IEEE.

[236] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2247–2256.

[237] Xuan Xie, Kristian Kersting, and Daniel Neider. 2022. Neuro-Symbolic Verification of Deep Neural Networks. In *Proc. of IJCAI-22*. https://doi.org/10.48550/ARXIV.2203.00938

[238] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. 2018. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. arXiv:1711.11157 [cs.AI]

[239] Ran Xu, Jayoung Lee, Pengcheng Wang, Saurabh Bagchi, Yin Li, and Somali Chaterji. 2022. LiteReconfig: cost and content aware reconfiguration of video object detection systems for mobile GPUs. In *Proceedings of the Seventeenth European Conference on Computer Systems* (Rennes, France) *(EuroSys '22)*. Association for Computing Machinery, New York, NY, USA, 334–351. https://doi.org/10.1145/3492321.3519577

[240] Ran Xu, Jayoung Lee, Pengcheng Wang, Saurabh Bagchi, Yin Li, and Somali Chaterji. 2022. LiteReconfig: cost and content aware reconfiguration of video object detection systems for mobile GPUs. In *Proceedings of the Seventeenth European Conference on Computer Systems* (Rennes, France) *(EuroSys '22)*. Association for Computing Machinery, New York, NY, USA, 334–351. https://doi.org/10.1145/3492321.3519577

[241] Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Fangzhou Mu, Yin Li, and Yingyu Liang. 2024. Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning. In *International Conference on Learning Representations*.

[242] Guitao Yang, Hamed Rezaee, and Thomas Parisini. 2019. Sensor redundancy for robustness in nonlinear state estimation. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 3865–3870.

[243] Tianci Yang, Carlos Murguia, Margreta Kuijper, and Dragan Nešić. 2020. A multi-observer based estimation framework for nonlinear systems under sensor attacks. *Automatica* 119 (2020), 109043.

[244] Yupeng Yang, Yiwei Lyu, and Wenhao Luo. 2023. Minimally constrained multi-robot coordination with line-of-sight connectivity maintenance. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7684–7690.

[245] Yupeng Yang, Yiwei Lyu, Yanze Zhang, Ian Gao, and Wenhao Luo. 2024. Integrating Online Learning and Connectivity Maintenance for Communication-Aware Multi-Robot Coordination. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5770–5776.

[246] Yupeng Yang, Yiwei Lyu, Yanze Zhang, Sha Yi, and Wenhao Luo. 2024. Decentralized Multi-Robot Line-of-Sight Connectivity Maintenance under Uncertainty. In *Proceedings of Robotics: Science and Systems*. Delft, Netherlands. https://doi.org/10.15607/RSS.2024.XX.005

[247] Shuochao Yao, Yiran Zhao, Huajie Shao, Chao Zhang, Aston Zhang, Shaohan Hu, Dongxin Liu, Shengzhong Liu, Lu Su, and Tarek Abdelzaher. 2018. Sensegan: Enabling deep learning for internet of things with a semi-supervised framework. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 3 (2018), 1–21.

[248] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2025. MiniCPM-v: A GPT-4v level MLLM on your phone. *Nature Communication (in press)* (2025).

[249] Ying C Yeh. 1996. Triple-triple redundant 777 primary flight computer. In *1996 IEEE Aerospace Applications Conference. Proceedings*, Vol. 1. IEEE, 293–307.

[250] Xiang Yin, Majid Zamani, and Siyuan Liu. 2021. On approximate opacity of cyber-physical systems. *IEEE Trans. Automat. Control* 66, 4 (2021), 1630–1645.

[251] Sze Zheng Yong, Minghui Zhu, and Emilio Frazzoli. 2018. Switching and data injection attacks on stochastic cyber-physical systems: Modeling, resilient estimation, and attack mitigation. *ACM Transactions on Cyber-Physical Systems* 2, 2 (2018), 1–2.

[252] Xinghuo Yu and Yusheng Xue. 2016. Smart grids: A cyber–physical systems perspective. *Proc. IEEE* 104, 5 (2016), 1058–1070.

[253] Yue Yu, Honglun Wang, and Na Li. 2019. Fault-tolerant control for over-actuated hypersonic reentry vehicle subject to multiple disturbances and actuator faults. *Aerospace Science and Technology* 87 (2019), 230–243.

[254] Zhenhua Yu, Hongxia Gao, Xuya Cong, Naiqi Wu, and Houbing Herbert Song. 2023. A survey on cyber–physical systems security. *IEEE Internet of Things Journal* 10, 24 (2023), 21670–21686.

[255] M. Zamani, P. Mohajerin Esfahani, R. Majumdar, A. Abate, and J. Lygeros. 2014. Symbolic control of stochastic systems via approximately bisimilar finite abstractions. *IEEE Transactions on Automatic Control, **Special Issue on Control of Cyber-Physical Systems*** 59, 12 (November 2014), 3135–3150.

[256] M. Zamani, G. Pola, M. Mazo Jr., and P. Tabuada. 2012. Symbolic models for nonlinear control systems without stability assumptions. *IEEE Transaction on Automatic Control* 57, 7 (July 2012), 1804–1809.

[257] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. 2024. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235* (2024).

[258] Jiangning Zhang, Haoyang He, Xuhai Chen, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. 2024. Gpt-4v-ad: Exploring grounding potential of vqa-oriented gpt-4v for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence*. Springer, 3–16.

[259] K. Zhang, X. Yin, and M. Zamani. 2019. Opacity of nondeterministic transition systems: A (bi)simulation relation approach. *IEEE Trans. Automat. Control* 64, 12 (2019), 5116–5123.

[260] Lin Zhang, Luis Burbano, Xin Chen, Alvaro A Cardenas, Steven Drager, Matthew Anderson, and Fanxin Kong. 2024. Fast Attack Recovery for Stochastic Cyber-Physical Systems. In *2024 IEEE 30th Real-Time and Embedded Technology and Applications Symposium (RTAS)*. IEEE, 280–293.

[261] Lin Zhang, Xin Chen, Fanxin Kong, and Alvaro A Cardenas. 2020. Real-time attack-recovery for cyber-physical systems using linear approximations. In *2020 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 205–217.

[262] Lin Zhang, Pengyuan Lu, Fanxin Kong, Xin Chen, Oleg Sokolsky, and Insup Lee. 2021. Real-time attack-recovery for cyber-physical systems using linear-quadratic regulator. *ACM Transactions on Embedded Computing Systems (TECS)* 20, 5s (2021), 1–24.

[263] Xumiao Zhang, Anlan Zhang, Jiachen Sun, Xiao Zhu, Y Ethan Guo, Feng Qian, and Z Morley Mao. 2021. Emp: Edge-assisted multi-vehicle perception. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (Mobicom)*. 545–558.

[264] Haonan Zhao, Yiting Wang, Thomas Bashford-Rogers, Valentina Donzella, and Kurt Debattista. 2024. Exploring generative AI for sim2real in driving data synthesis. In *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 3071–3077.

[265] Pai Zheng, Honghui Wang, Zhiqian Sang, Ray Y Zhong, Yongkui Liu, Chao Liu, Khamdi Mubarok, Shiqiang Yu, and Xun Xu. 2018. Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, and future perspectives. *Frontiers of Mechanical Engineering* 13 (2018), 137–150.

[266] Xugui Zhou, Bulbul Ahmed, James H. Aylor, Philip Asare, and Homa Alemzadeh. 2021. Data-driven Design of Context-aware Monitors for Hazard Prediction in Artificial Pancreas Systems. In *51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. 484–496. https://doi.org/10.1109/DSN48987.2021.00058

[267] Xugui Zhou, Bulbul Ahmed, James H Aylor, Philip Asare, and Homa Alemzadeh. 2023. Hybrid knowledge and data driven synthesis of runtime monitors for cyber-physical systems. *IEEE Transactions on Dependable and Secure Computing* 21, 1 (2023), 12–30.

[268] Xugui Zhou, Anqi Chen, Maxfield Kouzel, Haotian Ren, Morgan McCarty, Cristina Nita-Rotaru, and Homa Alemzadeh. 2025. Runtime Stealthy Perception Attacks against DNN-Based Adaptive Cruise Control Systems. In *ACM Asia Conference on Computer and Communications Security (ASIA CCS)*.

[269] Xugui Zhou, Maxfield Kouzel, and Homa Alemzadeh. 2022. Robustness testing of data and knowledge driven anomaly detection in cyber-physical systems. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE, 44–51.

[270] Xugui Zhou, Anna Schmedding, Haotian Ren, Lishan Yang, Philip Schowitz, Evgenia Smirni, and Homa Alemzadeh. 2022. Strategic safety-critical attacks against an advanced driver assistance system. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 79–87.

[271] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of The 7th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 229)*, Jie Tan, Marc Toussaint, and Kourosh Darvish (Eds.). PMLR, 2165–2183.

Received XXX; revised XXX; accepted XXX